Lab 2 Final Report: COVID-19 Case Rate vs Population Demographics and Mask Policy

Aidan Jackson, Frank Liu, Sam Temlock, Haoyu Zhang

Initial reassignment of common variables used across models:

```
df <- read.csv("covid-19.csv", header = TRUE)
df<-df%>%
  rename(case_rate_100k = 'Case.Rate.per.100000',
         population_density = 'Population.density.per.square.miles',
         mask_public='Mandate.face.mask.use.by.all.individuals.in.public.spaces',
         poverty_pct = "Percent.living.under.the.federal.poverty.line..2018.",
         unemployed_pct = "Percent.Unemployed..2018.",
         senior_pct = "X65."
         ) %>%
  select(State, case_rate_100k, population_density, mask_public, poverty_pct,
         unemployed_pct, senior_pct)
```

1. An Introduction

Research Question: How is the COVID-19 case rate related to the distribution of population demographics and policy within a state?

For this report, the investigation will be centered around the relationship of population demographics and policy on the COVID-19 case rates across the United States (US), which is grouped by the 50 states plus the District of Columbia (D.C.). Within the report, this collection will be referred to as the "states", and each member a "state", inclusive of D.C..

The research question aims to analyze the relationship between the COVID-19 case rate per 100,000 residents within each state (recorded between January 21, 2020 and October 30, 2020) and the population demographics features, as well as the policy decisions that were or were not put into place in order to combat the rise of COVID-19 cases. In this sense, the aim is to examine how dependent the COVID-19 case rate is on that of which cannot be controlled (i.e., population demographics), as well as that of which can be controlled to a certain degree (i.e., implementation of policies to attempt to combat case rate) by each state. With this information, one could suggest whether or not the proliferation of COVID-19 within a state seems to be related to either or both controllable and uncontrollable factors, and how they differ. The modeling goal of this research question will be one of description, and will be broken up into three phases of investigation.

The first and primary phase of investigation involves how the COVID-19 case rate is related to key population demographics. For this, the key variables are COVID-19 case rate per 100,000 residents and the senior rate (defined as 65 years old or older in the original dataset) within a state. The case rate per 100,000 was selected as the key output/dependent variable as it provides a standardized measure of the spread of COVID-19 across the states, and already takes into account the absolute population of the states. The senior percentage was selected as the key input/independent variable given that guidance had been released by the Centers for Disease Control and Prevention (CDC) that seniors belong to the age category for those at higher risk of COVID-19, and thus are a well studied and documented group. This is likely due to both the fact that seniors

in this age group are more likely or had greater access to be tested given their categorization, and the fact that they would have a greater likelihood of exhibiting detectable COVID-19 related symptoms that would prompt them to get tested. An additional rationale for measuring the senior percentage is that they may also be more likely to contract COVID-19 at lower viral loads, making them more susceptible to the virus. Conversely, this may result in seniors being more wary of the threat of COVID-19, and taking additional precautions to prevent infection relative to other age categories, such as limiting social interactions and taking more preventative measures in terms of hygiene. These key variables will provide an initial understanding of the relationship between the spread of the virus and population demographics.

Following this, in the second phase of investigation the analysis considers other variables of state-level population demographic features that are considered factors of susceptibility. Specifically, these features are the rate of poverty (defined as the percent of individuals living under the federal poverty line in 2018 in the original dataset) and the rate of unemployment (defined as the percent unemployment in 2018 in the original dataset). Those living below the poverty line may have less access/are unable to afford preventative controls such as sanitation products and masks that help prevent the spread of the virus, while those who are designated as unemployed may not have the flexibility of sheltering at home nor have access to the aforementioned preventative controls. Alternatively, the likelihood of reduced mobility for those living under the poverty line and/or those who are unemployed may also play a role in the relationship with case rate. As mentioned, these variables operationalize population metrics that may lead to greater case rates, and are already standardized as rates to account for varying absolute populations across states. It should be qualified that the data for poverty rate and unemployment rate are from 2018, and therefore may not provide a fully accurate reflection of the respective rates at the time that the case rate data was recorded.

The tertiary and final phase of investigation includes a variable that measures the implementation of policy as a response to COVID-19, and to understand the added effect of this variable in conjunction with the other phases of investigation. Specifically, there is a focus on a policy that mandates the wearing of masks in public spaces within the state. Through this, the aim is to analyze whether the implementation of a mask-related policy is related to case rate, as well as the strength of the association relative to that of population metrics. To operationalize this policy and simplify the measurements, the model will only consider whether or not this policy was implemented through a transformed indicator variable (1 = implemented, 0 = not implemented). As a result, factors of when and for how long the policy was implemented will be lost. Although this loss of information may fail to capture the impact of length of time of a policy on case rate (the policy may take time to be adopted/show meaningful efficacy), as there is no time-series data for case rate included in the data set, it was adjudged to be incongruent with the analysis.

Additionally, the final phase of investigation continues to expand on the examination of state-level population demographics via the population density variable (defined as the population density in square miles, population/square miles in the original dataset). A key factor to the spread of COVID-19 is the idea of social distancing, where the virus is considered to be more likely to spread when people are close in proximity. Thus, population density is measured as a conduit to indicate the level of proximity within each state and will operationalize the concept of social distancing. It should be noted however that the population density data may be limited in its ability to serve as a direct measurement of social distancing given it lacks other contextualize information, such as the urban to rural ratio as later discussed in section 5.

1-1. Assumptions

Prior to the analysis, it is important to identify a set of assumptions that have been made throughout the report and to assess the appropriateness of the data. Although there may be other considerations against the appropriateness of the data, the following highlights three particular arguments that must be taken into account when interpreting the results of the analysis.

Firstly, it is important to note that given each state is treated as a unique data point, the sample contains 51 data points. Although this size meets the general rule that an adequate sample size is 30 data points or more, as the analysis begins to factor in the wide range of potential population distributions and demographics within the states, it is clear that there is large variation within the sample. For example, the population

density can have large variability depending on the population concentration among a few large areas as well as the level of uninhabited or sparsely inhabited land within a state, while the demographics of population can depend on a wide range of things such as regional factors and employment opportunities. This point is further discussed in the IID assumption addressed in section 3 of this report.

Secondly, note that there are many internal and external aspects of the selected variables that have not been included within the models. Just addressing the internal information that is lost, it can be seen that some of the information is not captured given the methods of operationalization discussed previously. Among others, the loss of information on the date of implementation within the mask policy variable strips out any contextual knowledge regarding the length of policy implementation, and precludes the identification of how long a specific policy was implemented. Assuming that this has some relationship with case rate, this difference in length of time may play a part. There are also other external aspects that cannot be included, such as the rural to urban ratio as opposed to population density.

Finally, with regard to the policy variables, this report focuses on a form of mask mandate. Given this, it does not capture other policies that may or may not have other relationships with the case rate (e.g., implementation of stay-at-home orders, closure of non-essential businesses, etc.). In justification, these choices were largely made due to there either being an appropriate amount of samples in each category (e.g., most states have implemented basic policies such as the closure of non-essential businesses and the implementation of stay-at-home orders, and thus these were not included given the lack of samples for the states that have not implemented them), or that data on other policies were simply not readily available in the dataset. The same argument can be applied in the selection of population demographics.

2. A Model Building Process

The primary variable of interest will be the total COVID-19 case rate per 100,000 residents. The COVID-19 case rate was judged to be able to provide a better understanding than the related death rate because of the potential for other, unrecorded variables that could be correlated with a person dying from COVID-19 versus just becoming infected. For example, the availability and quality of medical care in a state may impact its ability to keep COVID-19 infected patients alive, and these variables are not included in the data set. The health status of the residents in one state compared to another state may also affect the death rate, but this is also not included. Instead, the COVID-19 case rate was chosen so that these other variables which may correlate with the death rate would not have to be considered.

The covariates that will be examined in building the models fit into two broad categories. The first category is the demographic features of the state, namely senior rate, the unemployment rate, the poverty rate, and the population density. The second category is that of policy decisions taken by the state, specifically a mask mandate. Problematic covariates would include any of the other direct measures of COVID-19 severity in the state, namely total infection rate not on a per capita basis and COVID-19 death rates. It can be assumed that these variables would be collinear with the primary variable of interest since COVID-19 case rate is simply a linear transformation of total cases not adjusted for population.

The three models aim to investigate the relationship between the COVID-19 case rate, state population demographics, and policy (i.e., mask mandate). COVID-19 case rate per 100,000 is chosen as the key dependent variable to represent the proliferation of the COVID-19 pandemic in each state. The population percentage of seniors is chosen as the key independent variable as a proxy for susceptibility by innate state population make-up. The variables of poverty percentage, unemployment percentage, mask mandate policy, and population density per square miles are chosen as candidate covariates to further explore the aforementioned potential relationship.

In order to allow for more simplistic and direct interpretations when analyzing the relationships of the input variables on the output, it was determined that all percentage variables (senior, poverty, and unemployment) would be transformed into rates per 100,000 to align with the case rate variable. It will also serve to ease the burden on the readers of this report as like-for-like scaled relationships can be drawn between the variables. Given that this transformation is constant for all values (scalar multiplication of 100,000 for the seniors

variable given it is a ratio, and 1,000 for poverty and unemployment given they are percentages on a scale of 0-100%), it will have no effect on the distributions of the variables.

```
# Percentage variables are re-scaled to rates per 100k
df <-df %>%
  mutate(
    senior_per_100k = 100000*senior_pct,
    poverty_per_100k = 1000*poverty_pct,
    unemployed_per_100k = 1000*unemployed_pct
  )
```

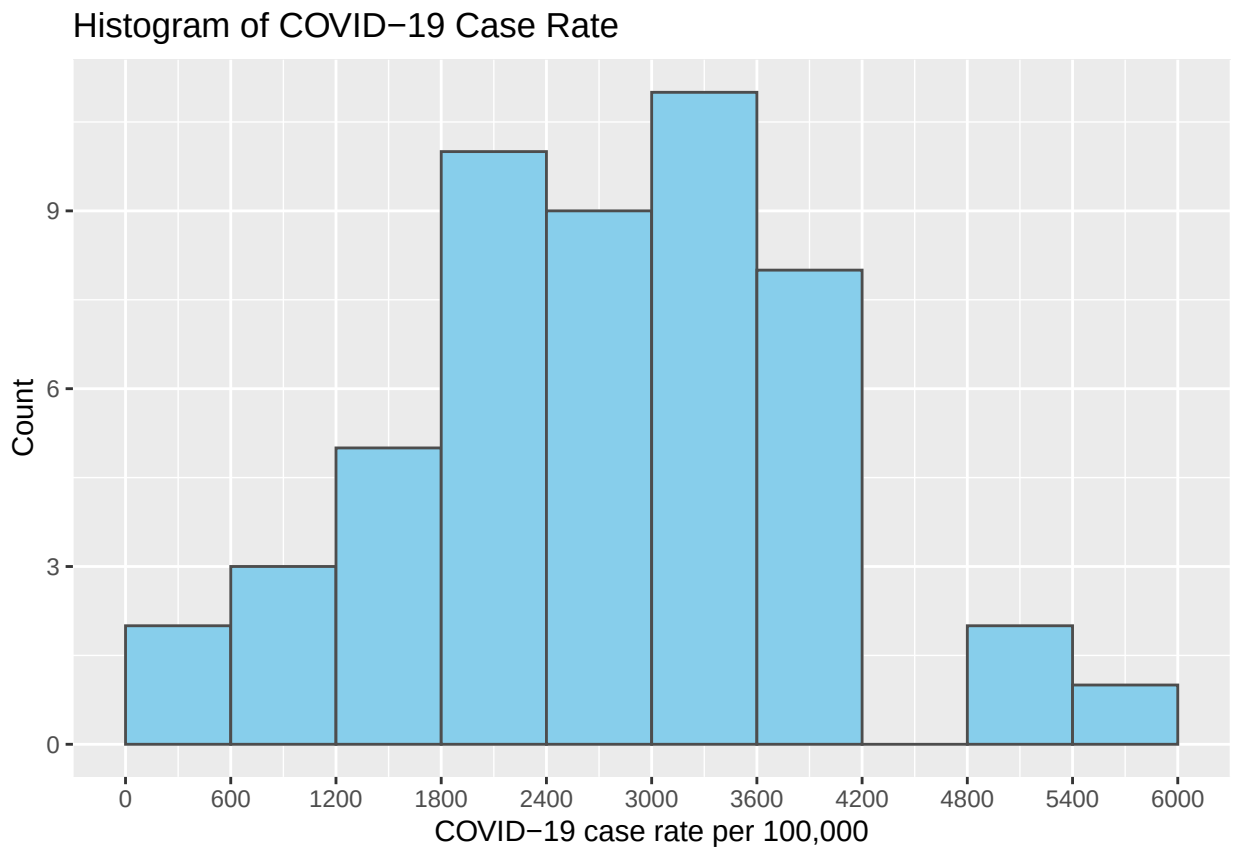
2-1. Model 1

2-1-1. Model 1 Exploratory Data Analysis

For the first model, the relationship between the COVID-19 case rate per 100,000 and senior rate was analyzed.

First, the distribution of the case rate dependent variable is examined.

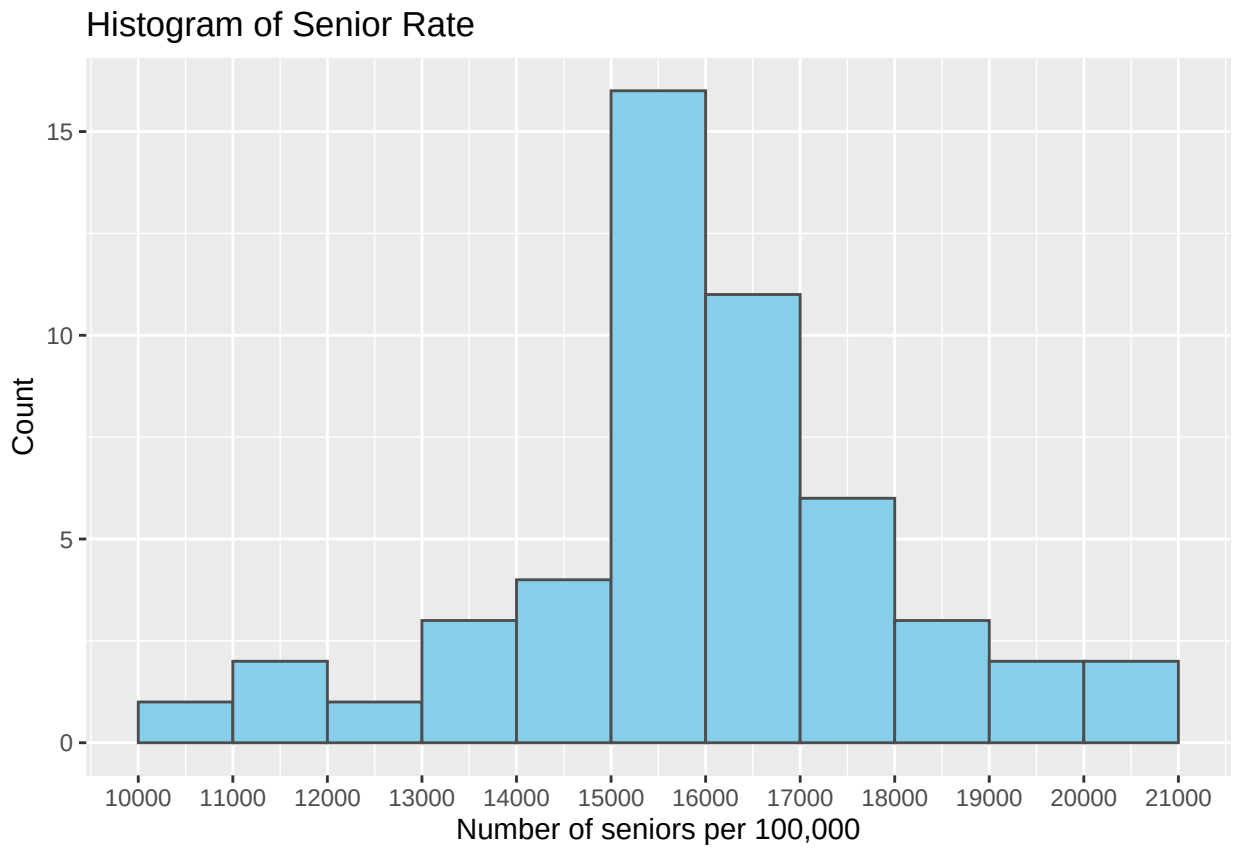
```
ggplot(data = df,
  mapping = aes(x= case_rate_100k)) +
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(0,6000,600)) +
  labs(title =  "Histogram of COVID-19 Case Rate",
       x = "COVID-19 case rate per 100,000", y = 'Count') +
  scale_x_continuous(breaks=seq(0, 6000, 600))
```



As can be seen above, the distribution is fairly normal, and given that it has already been standardized as a rate across all states, there is no need to perform any transformations on this variable. Thus, the case rate per 100,000 variable can be leveraged as is as the dependent variable for all three models.

Next, the distribution of the senior percentage transformed into the rate per 100,000 is examined.

```
ggplot(data = df,
  mapping = aes(x= senior_per_100k)) +
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(10000,21000,1000)) +
  labs(title = 'Histogram of Senior Rate',
  x = 'Number of seniors per 100,000', y = 'Count') +
  scale_x_continuous(breaks=seq(10000,22000,1000))
```

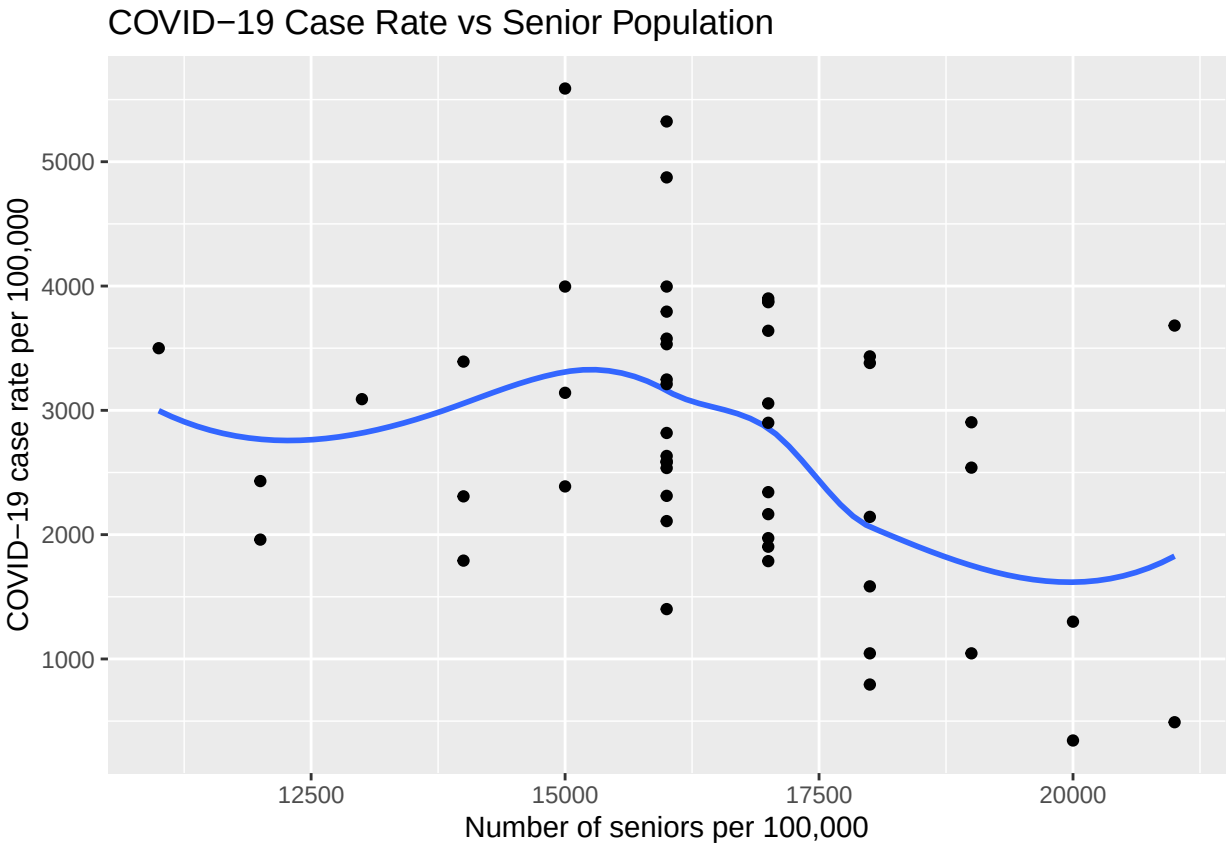


As can be seen from the distribution above, again there is a fairly normal distribution in the senior rate per 100,000 across the state. Additionally, there do not seem to be any outliers within the distribution, given that the range of rates fall between 11,000 and 21,000. Therefore, there are no additional transformations that need to be made in order to restructure the distribution, and this variable can be used across all three models.

With the appropriate variable distributions investigated, a plot is created to analyze the relationship between them.

```
df %>%
  ggplot(aes(senior_per_100k, case_rate_100k)) +
  geom_smooth(se = FALSE) +
  geom_point() +
  labs(
```

```
    title = 'COVID-19 Case Rate vs Senior Population',
    x = 'Number of seniors per 100,000',
    y = 'COVID-19 case rate per 100,000'
  )
```



From the above plot, there is noticeable variation in the smoothed blue line. However, it can be posited that there is a inverse relationship between the variables given the roughly linear relationship in the overall downward trend in the line.

2-1-2. Model 1 Regression

In order to test the relationship between the variables, the following equation is used to create a regression model for the case rate and senior rate variables.

$$case\_rate\_100k = \beta_0 + \beta_1 senior\_per\_100k$$

```
# Build the regression model for Model 1
model1 <- lm(case_rate_100k ~ senior_per_100k , data = df)
```

To improve the precision of the t-test of coefficients, classical standard errors can be used over robust standard errors to test the regression. To determine the applicability of using classical standard errors, in addition to the three Classical Linear Model (CLM) assumptions necessary for the robust standard errors, the CLM assumption for homoskedastic conditional errors should be evaluated. If this additional assumption is met, the robust standard errors can be replaced by the classical standard errors. The assumption of

homoskedasticity is met if the data does not have large variance among the residuals. The analysis for this assumption, as well as the other CLM assumptions, is conducted in subsection 2-1-3, titled "Model 1 Limitations".

Below, the t-test of coefficients is run on the regression model above.

```
coeftest(model1)
```

```
##
## t test of coefficients:
##
##                     Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     5613.666667 1233.338521  4.5516 3.531e-05 ***
## senior_per_100k   -0.173900    0.074307 -2.3403   0.02338 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

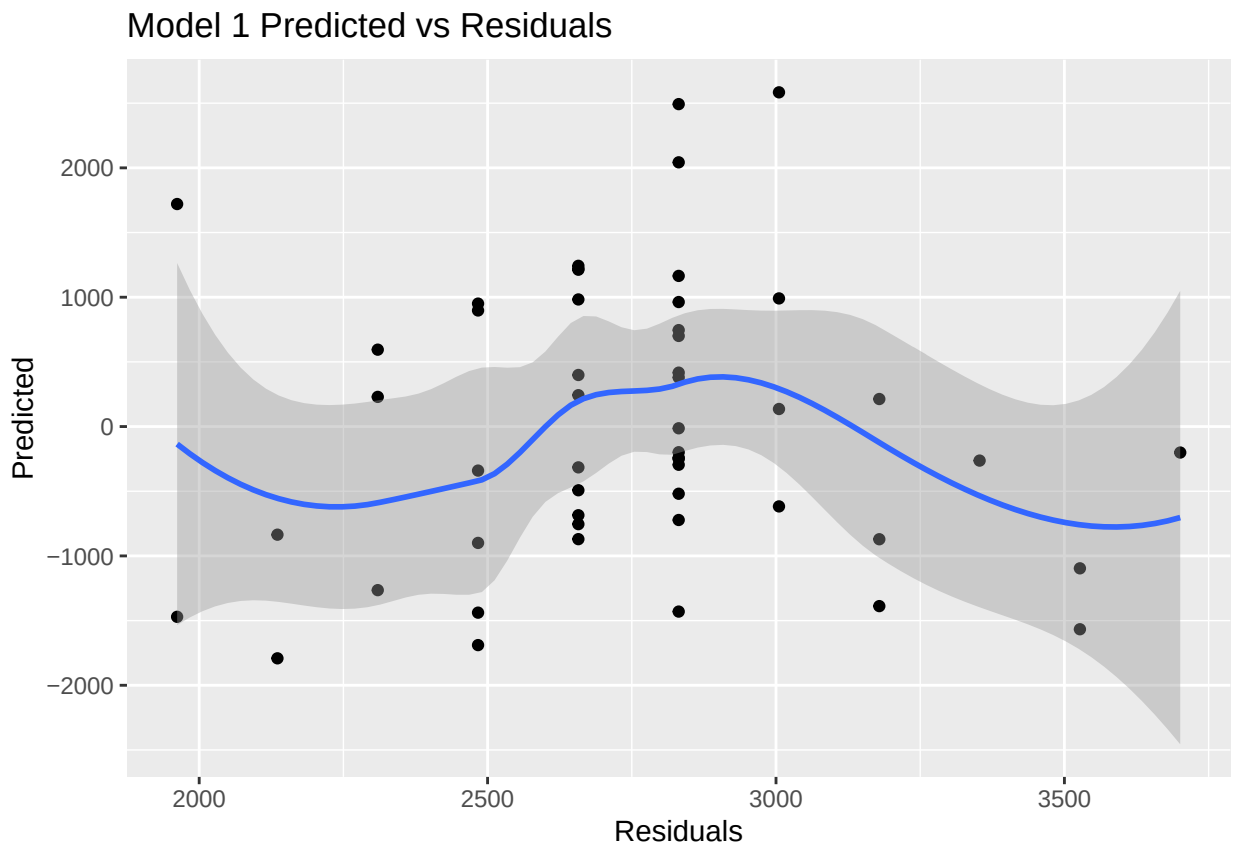
As can be seen from the above output, the variable for senior rate is shown to be significant within the model specification with a p-value of 0.023. Furthermore, the coefficient of -0.174 for the variable is negative, supporting the earlier evaluation that there is an inverse relationship between the senior rate and the case rate. In terms of practical significance, this can be interpreted as for every additional senior per 100,000 in the population, there is a decrease of 0.174 COVID-19 cases per 100,000. More intuitively, for roughly every 6 additional seniors per 100,000, there is a decrease of 1 case per 100,000.

2-1-3. Model 1 Limitations

i. IID Sampling The first assumption of IID is not detailed within this section given that it is evaluated more generally for all three models within section 3.

ii. Linear Conditional Expectation The second CLM assumption to be evaluated is that of a linear conditional expectation relationship within the model, to diagnose whether the model is accurately estimating the relationship between the variables. To test for this, the residuals of model 1 are plotted against the fitted values.

```
df<- df%>%
  mutate(
    model1_preds = predict(model1),
    model1_resids = resid(model1)
  )
df %>%
  ggplot(aes(model1_preds, model1_resids)) +
  geom_point() +
  stat_smooth() +
  labs(
    title = 'Model 1 Predicted vs Residuals',
    x = 'Residuals',
    y = 'Predicted'
  )
```

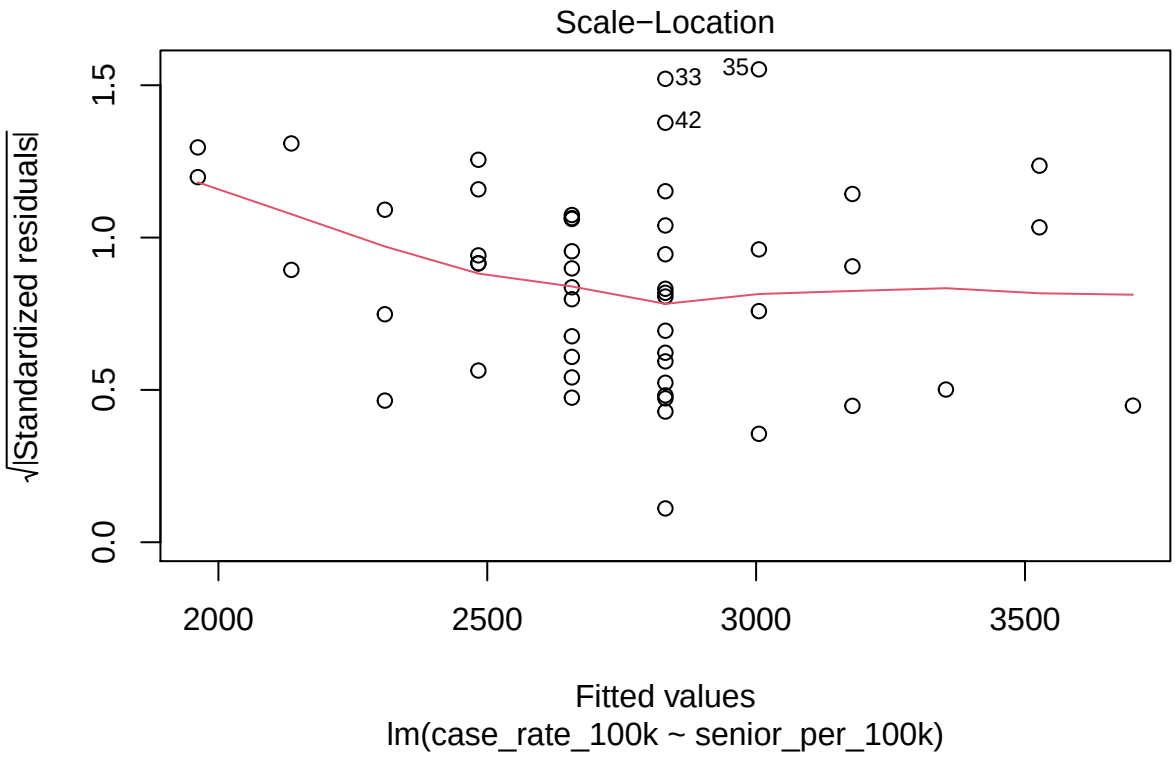


When examining the plot above for linearity in the conditional expectation, there is a noticeably slightly quadratic shape to the conditional expectation. This may suggest that the relationship between the variables may not be as linear as the model assumes. However, as discussed previously in the scatter plot of the two variables, the relationship can be approximated as roughly linear. It is also difficult to explain the complexity behind the COVID-19 case rate with just a single variable that does not fully capture the broader population demographics, so there is hope that given the addition of other input variables within models 2 and 3, the relationship within the data will be better explained and result in more linearity.

iii. No Perfect Collinearity There is no test conducted for the third CLM assumption of no perfect collinearity in model 1 given there is only one input variable.

iv. Homoskedastic Errors To evaluate the fourth CLM assumption of homoskedastic errors, the square root of the residuals (to remove any negative values) is plotted against the fitted values of the model. Additionally, a Breusch-Pagan test is run to check the level of heteroskedasticity.

```
# Check the homoskedasticity of errors in model 1
plot(model1, which=3)
```




```
# Run the Breusch-Pagan test for model 1
bptest(model1)
```

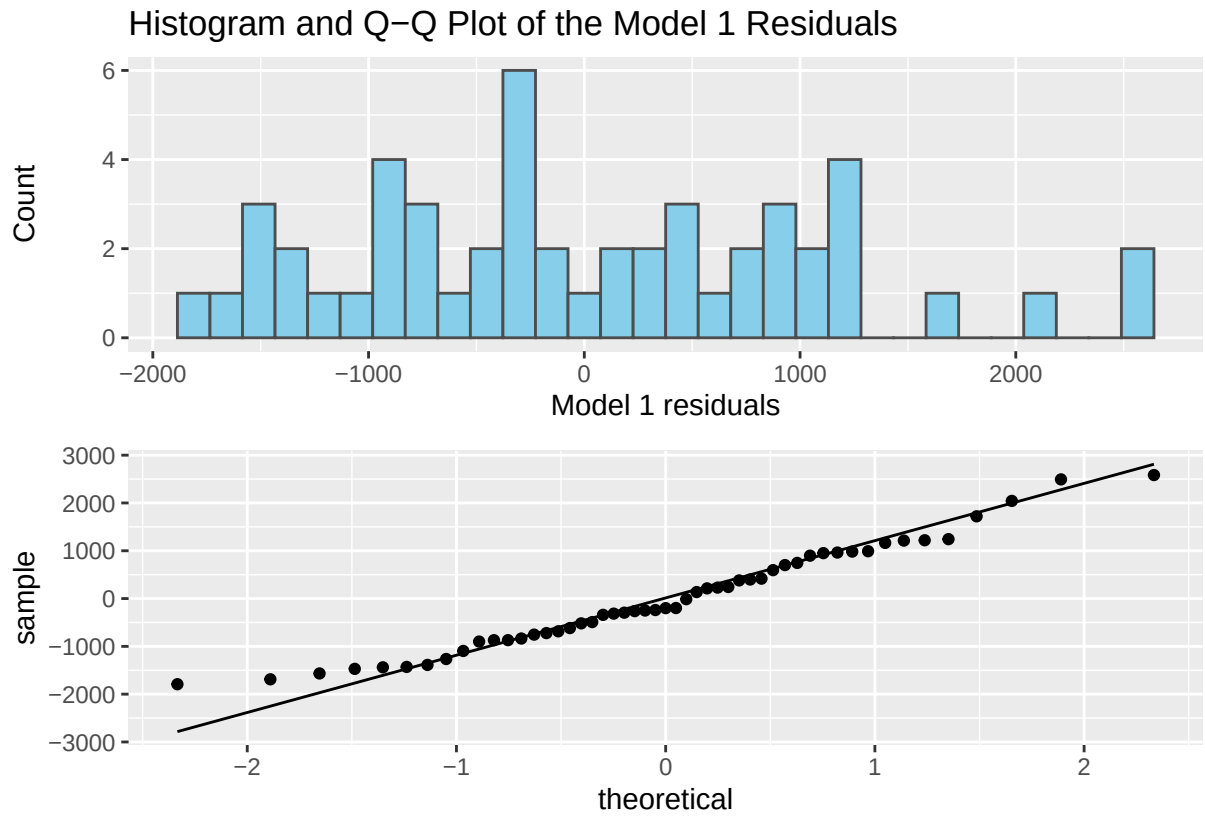
```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 0.46099, df = 1, p-value = 0.4972
```

From the above plot, the variance in residuals can be seen to be roughly constant given the relative flatness of the plotted red line, thus meeting the assumption of homoskedastic conditional errors. This is supported by the p-value of 0.497 from the Breusch-Pagan test, meaning it fails to reject the null hypothesis that the conditional errors are not heteroskedastic. For reference, the marked data points are New York (33), North Dakota (35), and South Dakota (42).

v. Normally Distributed Errors The final CLM assumption examined for model 1 is the assumption of normality of errors. This is a necessary assumption to ascertain that the errors used in the regression are drawn from a normal distribution, so that they can be accepted when used to calculate the significance levels. Both a Q-Q plot and histogram of the distribution of the residuals is plotted to observe the normality.

```
# Create a histogram of the distribution of the model 1 residuals
plot_one <- df %>%
  ggplot(aes(x = resid(model1))) +
  stat_bin() +
```

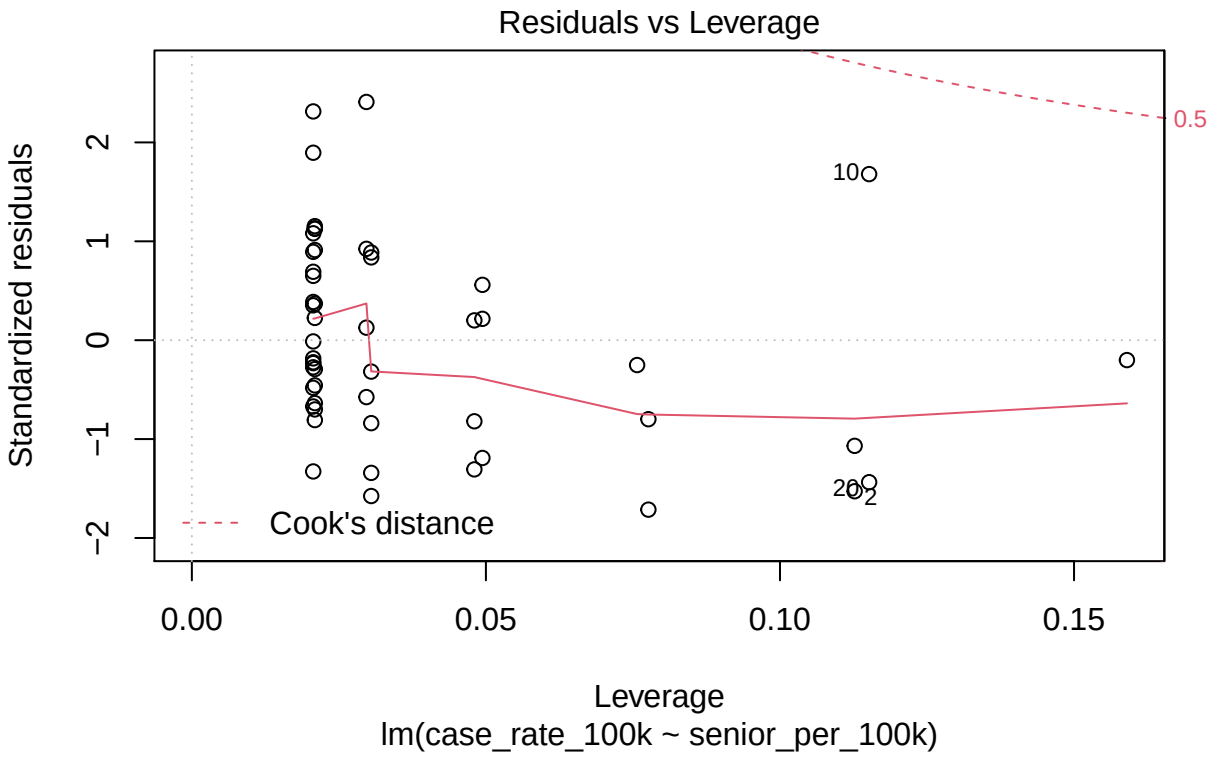
```
  geom_histogram(fill = 'skyblue', color = 'grey30', bins=30) +
  labs(title =  "Histogram and Q-Q Plot of the Model 1 Residuals",
       x = "Model 1 residuals", y = 'Count')

# Create a QQ plot for the model 1 residuals
plot_two <- df %>%
  ggplot(aes(sample = resid(model1))) +
  stat_qq() + stat_qq_line()

plot_one / plot_two
```



From the Q-Q plot, normality in the residuals can be detected given the proximity of the points to the normal line. Additionally, although the shape is a little harder to observe, the histogram of the residuals gives a fairly normal distribution.

vi. Influence of Data Points In addition to the CLM assumptions, the Cook's distance is also examined by a plot of the residuals vs the leverage to estimate the influence of the data points and identify any outliers in the values. From the plot below, no obvious outliers are detected. For reference, the marked data points are Alaska (2), Florida (10), and Maine (20).

```
# Investigate outliers using Cook's distance
plot(model1, which=5)
```



2-2. Model 2

2-2-1. Model 2 Exploratory Data Analysis

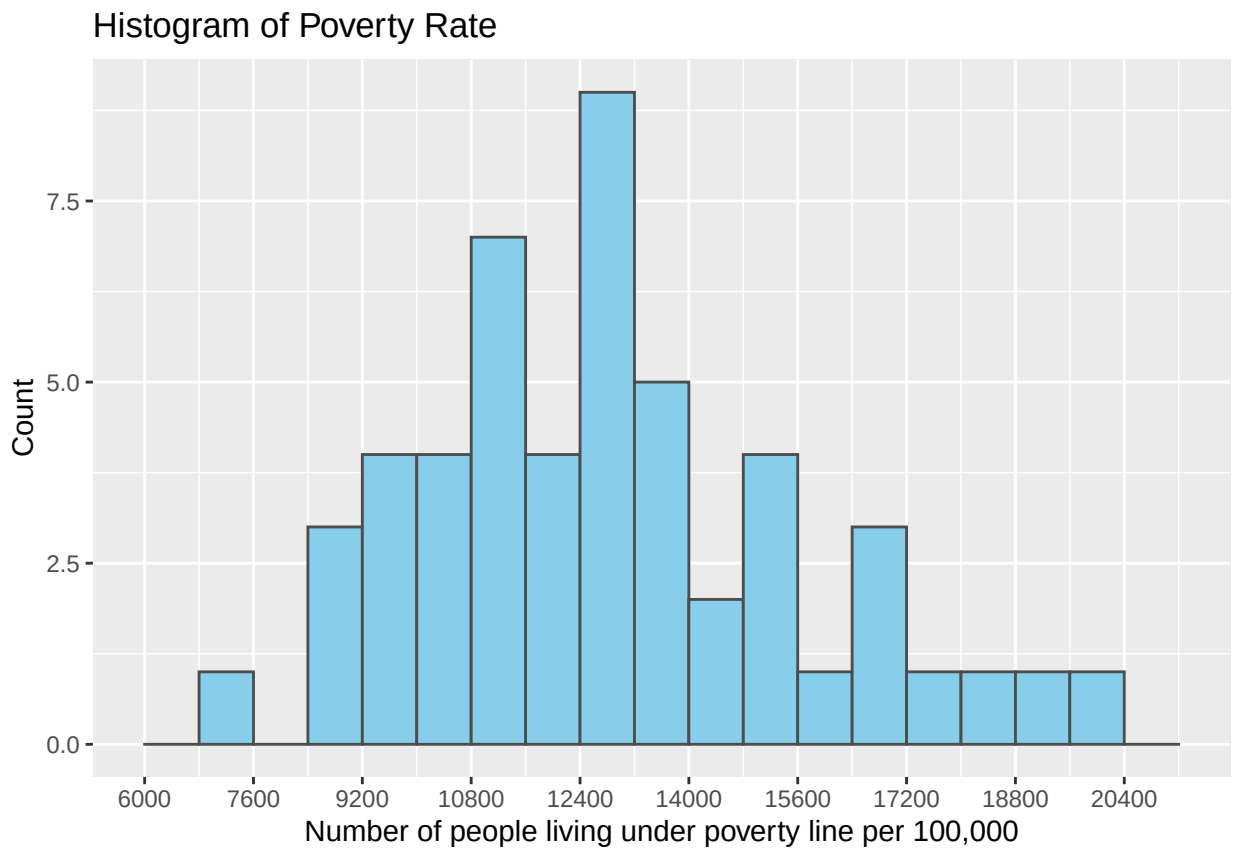
Building upon the work of model 1, two extra state demographic features regarding poverty rate (percentage living under the federal poverty line in 2018) and unemployment rate (percentage of unemployed in 2018) are included in the analysis of model 2. Although the increasing unemployment rate during the pandemic has been widely covered by the media, the potential pandemic causing poverty or unemployment will not be investigated here, according to the designed research question in this work.

Due to the fact that both poverty rate and unemployment rate data represents the status in 2018, it is proper to assume that they represent the pre-existing state demographic characteristics instead of the capturing any influence from the pandemic.

In addition, both newly-introduced percentage variables would be transformed into rates per 100,000 to align with the case rate variable.

First, the distribution of the poverty rate is examined as below.

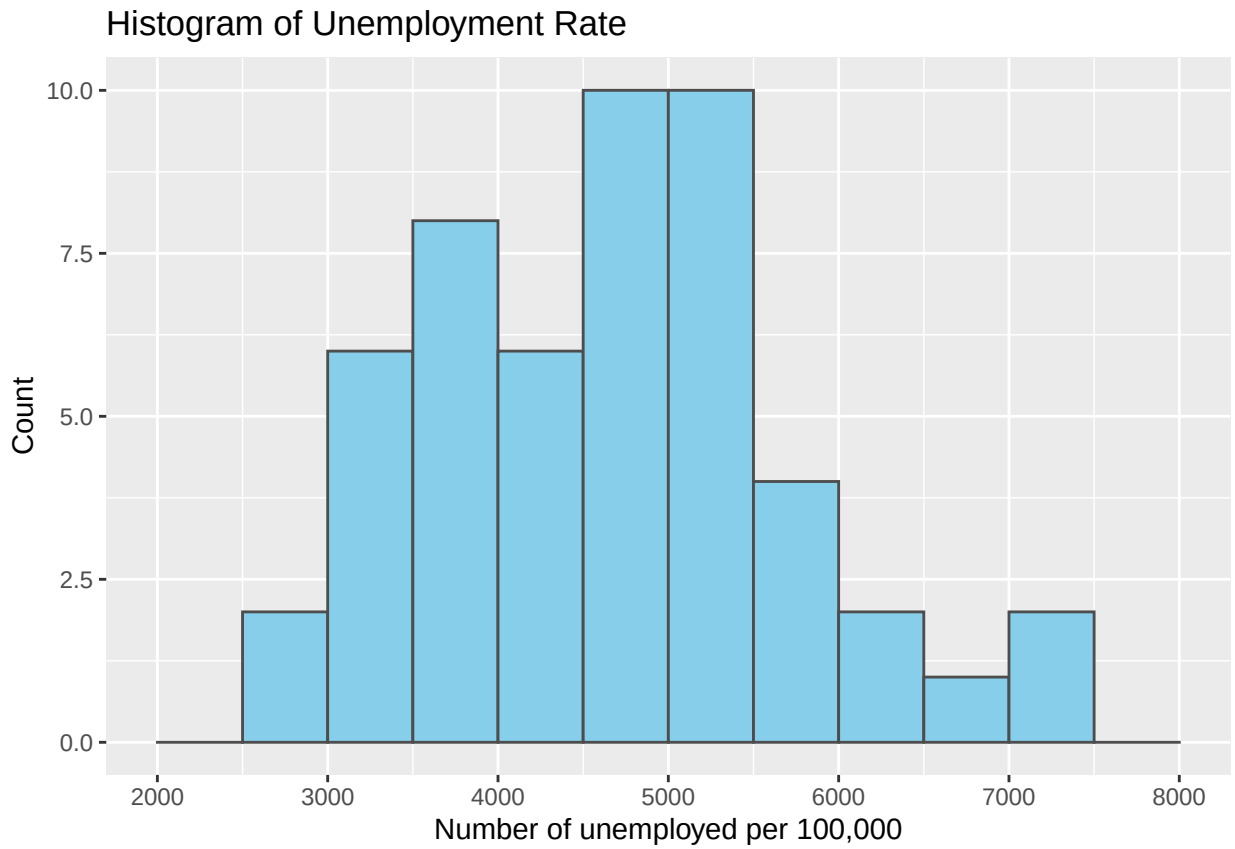
```
ggplot(data = df,
  mapping = aes(x= poverty_per_100k))+
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(6000,21200,800)) +
  labs(title =  "Histogram of Poverty Rate",
       x = "Number of people living under poverty line per 100,000", y = 'Count')+
  scale_x_continuous(breaks=seq(6000, 21200, 1600))
```



As shown in the above distribution, there is a fairly normal distribution in the poverty number per 100,000 across the states. Additionally, there is no outlier observed within the distribution, given that the range of rates fall between 7,600 and 19,700. Therefore, there is no further transformation needed for this variable in the following analysis process for both model 2 and model 3.

Next, the distribution of the unemployment rate is examined as below.

```
ggplot(data = df,
  mapping = aes(x= unemployed_per_100k))+
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(2000,8000,500)) +
  labs(title =  "Histogram of Unemployment Rate",
       x = "Number of unemployed per 100,000", y = 'Count')+
  scale_x_continuous(breaks=seq(2000, 8000, 1000))
```



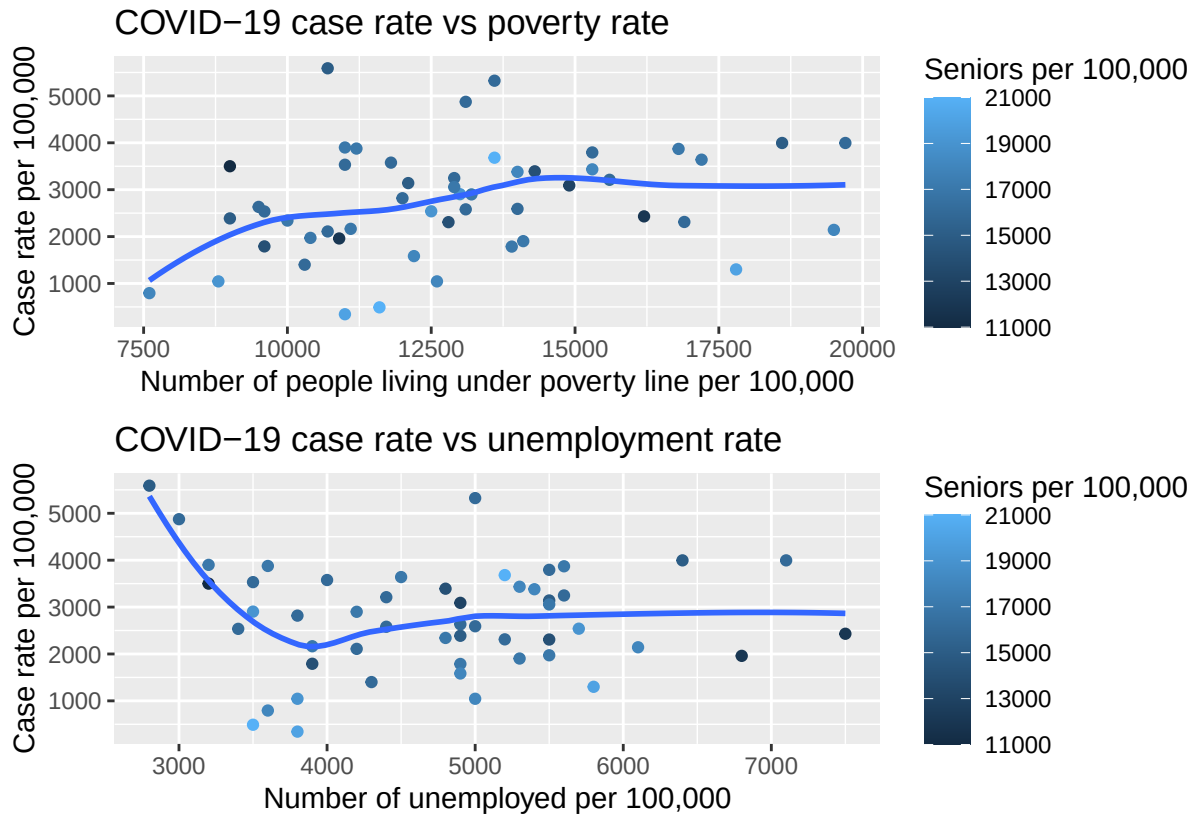
According to the histogram above, the distribution of the unemployed number per 100,000 is fairly normal. Overall, it is not highly skewed or heavily tailed. Therefore, there is no further transformation needed for this variable in the following analysis process for both model 2 and model 3.

2-2-2. Model 2 Regression

Compared to model 1, the relationship investigated here involves a higher dimensional space. It would be challenging to observe notable variation from the variable distribution plots as below.

```
case_poverty <- df %>%
  ggplot(aes(poverty_per_100k, case_rate_100k, color = senior_per_100k)) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'COVID-19 case rate vs poverty rate',
    x = 'Number of people living under poverty line per 100,000',
    y = 'Case rate per 100,000',
    color = 'Seniors per 100,000'
  )

case_unemployed <- df %>%
  ggplot(aes(unemployed_per_100k, case_rate_100k, color = senior_per_100k)) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'COVID-19 case rate vs unemployment rate',
```

```
    x = 'Number of unemployed per 100,000',
    y = 'Case rate per 100,000',
    color = 'Seniors per 100,000'
  )

case_poverty / case_unemployed
```



However, there is an overall increasing trend of COVID-19 case rate with increasing poverty rate, while there is an inverse relationship between the variables of the COVID-19 case rate and the unemployment rate.

Based on the results of the data exploration above, the following equation is used to create a regression model to determine the relationship between the case rate and demographic features including the senior rate, the rate of poverty and the rate of unemployment.

$$case\_rate\_100k = \beta_0 + \beta_1 senior\_per\_100k + \beta_2 poverty\_per\_100k + \beta_3 unemployed\_per\_100k$$

```
model2 <- lm(case_rate_100k ~ senior_per_100k + poverty_per_100k + unemployed_per_100k, data = df)
```

According to the assessment of homoskedasticity of model 2, which would be discussed in the following subsection (Model 2 Limitations - iv.Homoskedastic Errors), it is plausible to apply the classical standard errors in the t-test of coefficients here.

```r
coeftest(model2)
```

```
##
## t test of coefficients:
##
##                         Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)         5869.329734 1362.850759  4.3067 8.371e-05 ***
## senior_per_100k       -0.230970    0.068570 -3.3684 0.0015170 **
## poverty_per_100k       0.231750    0.062345  3.7172 0.0005355 ***
## unemployed_per_100k   -0.486191    0.168464 -2.8860 0.0058754 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the test results of the estimated coefficients in model 2, all three demographic features discussed here are significantly related to the COVID-19 case rate, especially for the poverty rate (p-value < 0.01).

The coefficient of the senior rate (senior_per_100k) is -0.230970 and is statistically significant, which implies that for every one thousand additional seniors per 100,000, there is a decrease of 231 COVID-19 cases per 100,000, holding other demographic features constant. As discussed in the subsection 2-1, although seniors are a susceptible group, their reduced mobility and social interactions with cautious self protection may play an important role in reducing the number of cases among seniors.

In addition, the coefficient of the poverty rate (poverty_per_100k) (0.231750) indicates that for every one thousand additional people living under the poverty line per 100,000, there is an increase of 232 COVID-19 cases per 100,000, holding other demographic features constant. People living under the poverty line may have limited access to medical resources. Moreover, they may not be able to afford to live in communities with appropriate sanitary conditions.

Finally, the coefficient of the unemployment rate (unemployed_per_100k) is statistically significant and negative (-0.486191), which implies that for every one thousand additional unemployed per 100,000, there is a decrease of 486 COVID-19 case per 100,000, holding other demographic features constant. Although unemployment leads to less medical insurance coverage and impacts income, the unemployed may have more flexibility to obey stay-at-home orders. In addition, there is less exposure probability given the unemployed do not commute to work or work in-person.

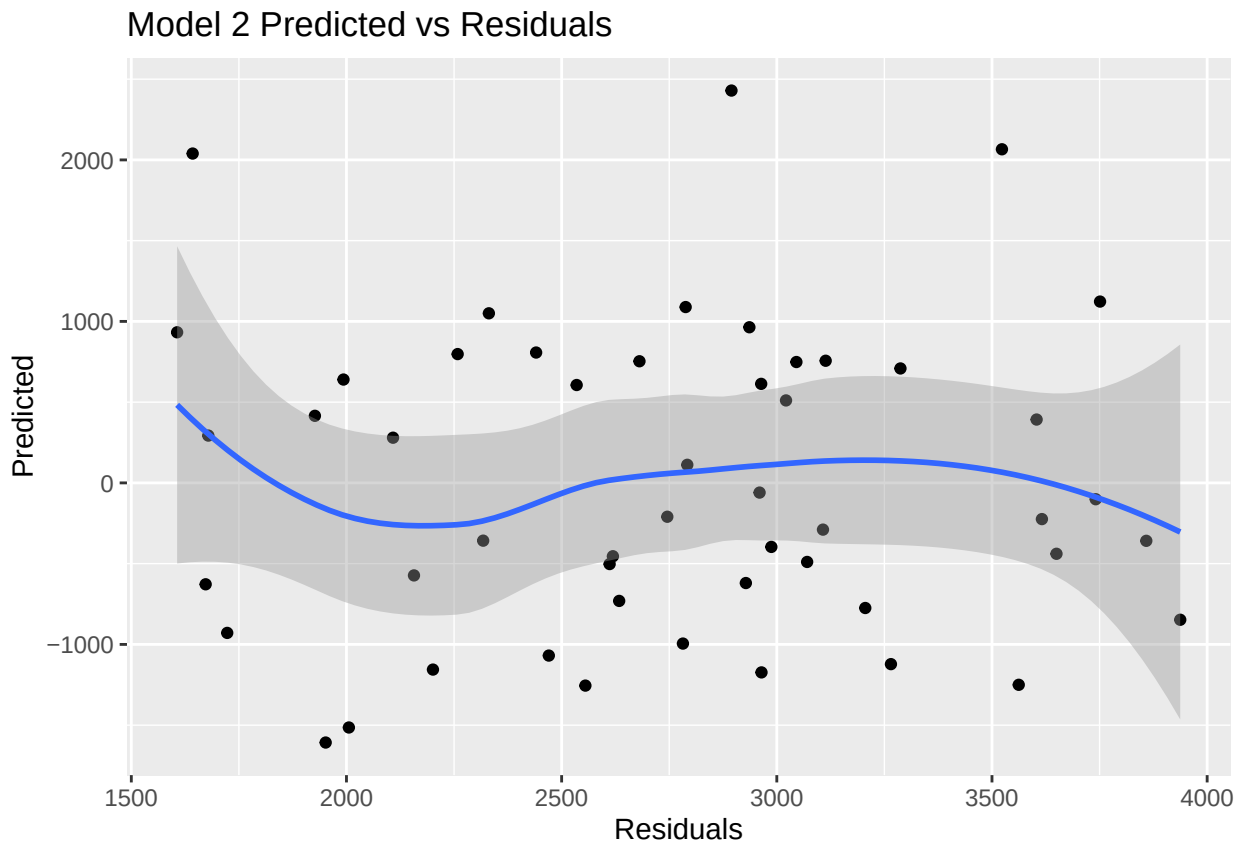
2-2-3. Model 2 Limitations

i. IID Sampling The first assumption of IID is not detailed within this section given that it is evaluated more generally for all three models within section 3.

ii. Linear Conditional Expectation To assess the linear conditional expectation assumption in a higher-dimensional space, the plot of the predicted values versus the residuals of model 2 is investigated as follows.

```r
df<- df%>%
  mutate(
    model2_preds = predict(model2),
    model2_resids = resid(model2)
  )
df %>%
  ggplot(aes(model2_preds, model2_resids)) +
  geom_point() +
  stat_smooth() +
```

```
  labs(
    title = 'Model 2 Predicted vs Residuals',
    x = 'Residuals',
    y = 'Predicted'
  )
```



Overall, the residuals remain around zero across the predicted range. There is no assumption violation phenomenon that can be observed here.

iii. No Perfect Collinearity Since there are multiple covariates included in this model compared to the model 1, there is concern about the no perfect collinearity assumption. First, the list of variable coefficients show that no variables were dropped, which means there was no perfect collinearity detected by the R function.

```
model2$coefficients
```

```
##         (Intercept)     senior_per_100k    poverty_per_100k unemployed_per_100k
##        5869.3297338          -0.2309704           0.2317495          -0.4861906
```

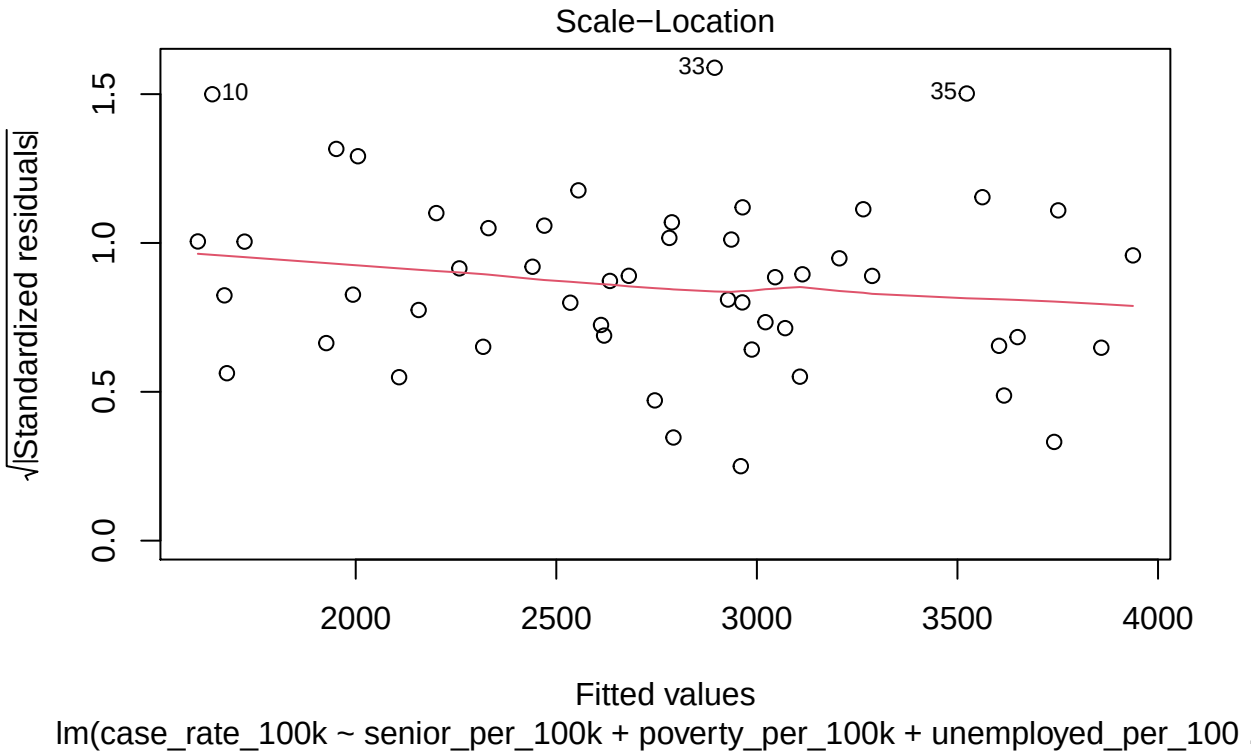
In addition, all the variance inflation factors are less than 4 as follows, which doesn't indicate the existence of high collinearity.

```
vif(model2)
```

```
##     senior_per_100k    poverty_per_100k unemployed_per_100k
##            1.065676           1.655600           1.692290
```

iv. Homoskedastic Errors To evaluate the homoskedastic errors assumption, the square root of the residuals is plotted against the fitted values of model 2 (scale-location plot). Although, the curve is not perfectly flat, there is no obvious variance in the errors either. For reference, the data points marked in the plot represent Florida (10), New York (33) and North Dakota (35).

```
plot(model2, which=3)
```



Additionally, a Breusch-Pagan test is run to check the level of heteroskedasticity. The test result shows it fails to reject the null hypothesis, which means there is no evidence for heteroskedasticity. Overall, the assumption of homoskedastic errors is satisfied for model 2.

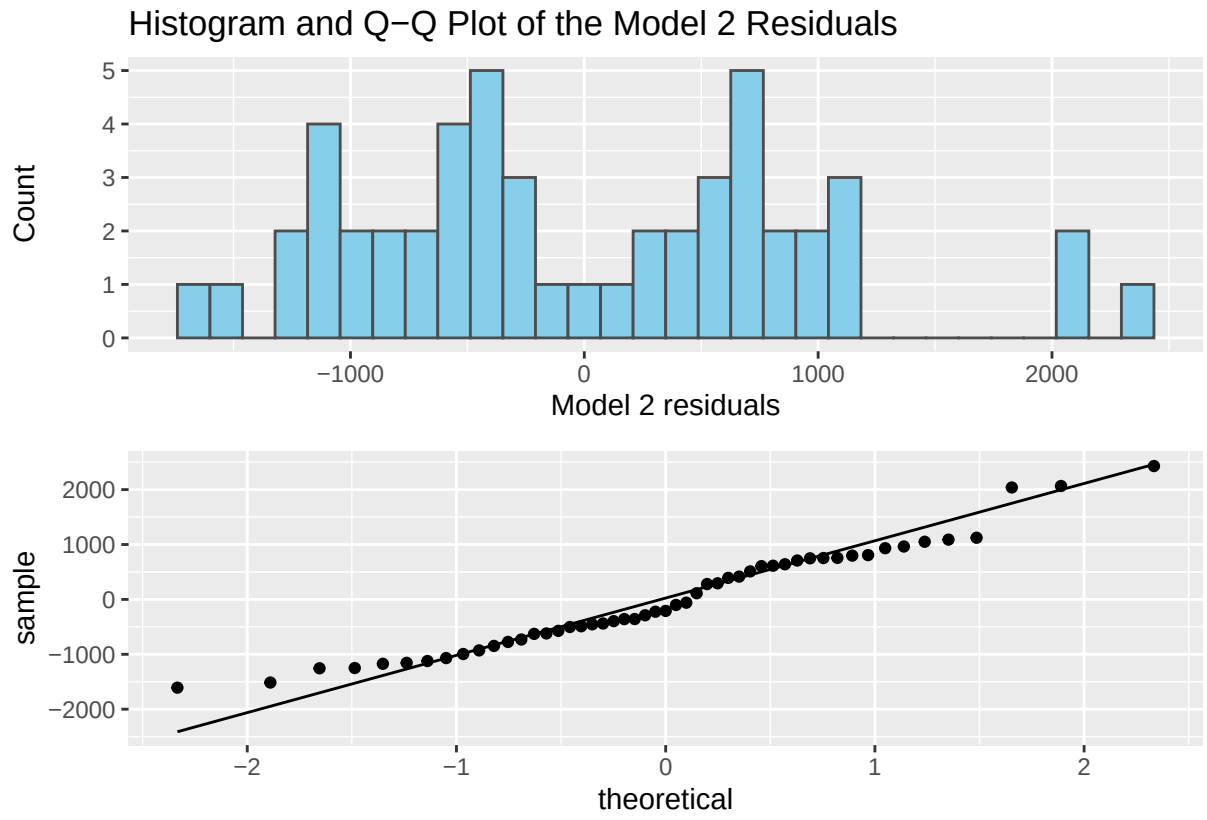
```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 4.5045, df = 3, p-value = 0.2119
```

v. Normally Distributed Errors To access the assumption of normality of error distribution, both the histogram and the Q-Q plot of the residuals are shown as follows.

```
plot_one <- df %>%
  ggplot(aes(x = model2_resids)) +
```

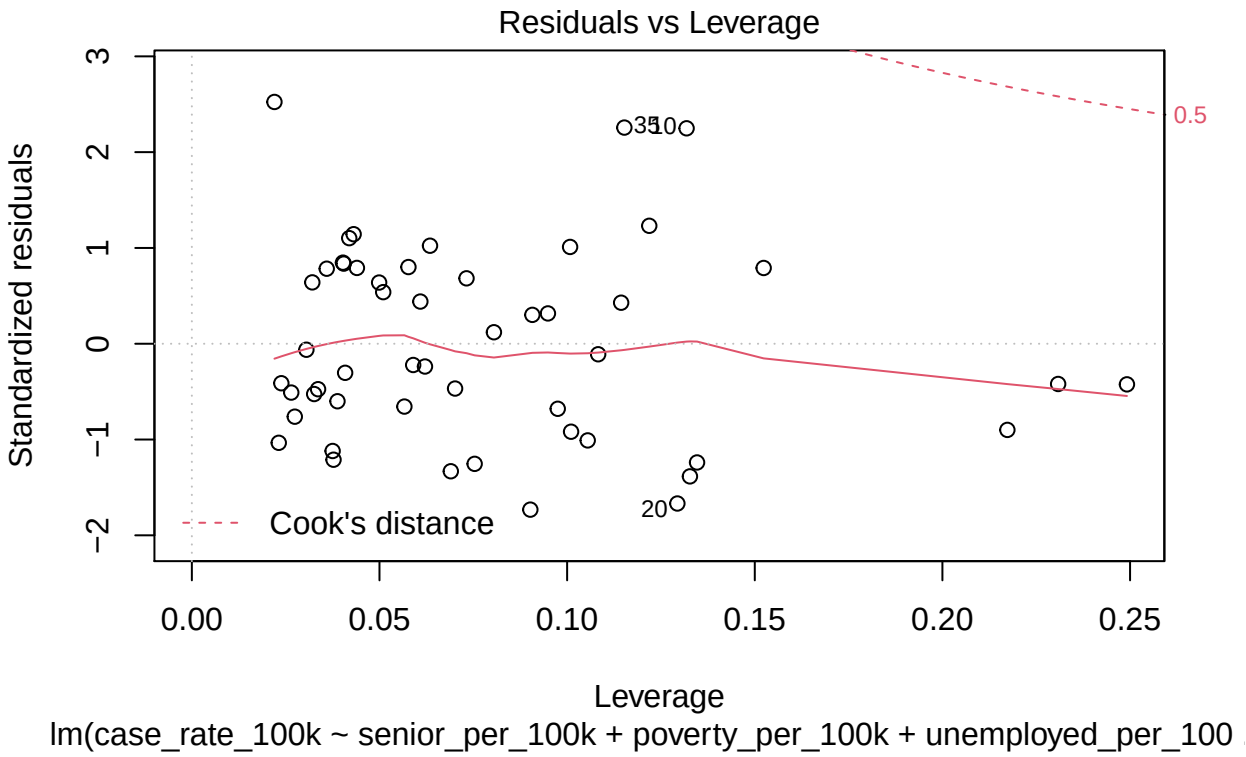
```
  geom_histogram(fill = 'skyblue', color = 'grey30', bins=30) +
  labs(title =  "Histogram and Q-Q Plot of the Model 2 Residuals",
       x = "Model 2 residuals", y = 'Count')

plot_two <- df %>%
  ggplot(aes(sample = model2_resids)) +
  stat_qq() + stat_qq_line()

plot_one / plot_two
```



Limited by the sample size, it is hard to observe a perfect normal distribution from the histogram. However, the Q-Q plot implies no deviation from normality. Overall, the normally distributed errors assumption is satisfied for model 2.

vi. Influence of Data Points Additionally, Cook's distance is used to estimate the influence of each single data point.

```
plot(model2, which = 5)
```



No obvious outliers could be detected here. For reference, the marked data points are Florida (10), Maine (20), and North Dakota (30).

2-3. Model 3

2-3-1. Model 3 Exploratory Data Analysis

For model 3, policy related variables will be introduced for the analytics. The policy variable "Mandate face mask use by all individuals in public spaces" (renamed as mask_public) is selected as an input variable in model 3. Given it has been proven in external research that masks are effective in preventing the transmission of COVID-19, it is expected that the mask policy mandate in public is related to a decrease in case rate.

There are other variables related to masks such as "No legal enforcement of face mask mandate" and "Mandate face mask use by employees in public-facing businesses". It is assumed these variables have a lower impact than the policy that is enforced to those in public. Also, it is observed that if a state enforces a public mask policy, they will also enforce a mask policy for employees in public-facing businesses. As a result, mask policy for all individuals in public spaces can be used to represent mask-related policies.

It is worth noting that the mask_public variable is in date format, which has either a value of 0 or an actual date when the policy was enforced. According to the documentation, 0 represents "the absence of an order or directive", which can be interpreted as the policy is not enforced by the state explicitly. For the linear regression, the date values in the mask_public variable are transformed into a value of 1, so that whether or not a state has a public mask policy can be distinguished. It is also noted that by transforming the variable, some important information will be lost because the actual date (early or late) of enforcement for the policy can also have an impact on case rate. However, it is difficult to measure the effect of the timing of implementation across different states. Additionally, the enforcement of a public mask policy will vary

across states depending on other state related features such as political factors that are not included within this analysis.

Firstly, the mask policy (mask_public_bool) variable is created from the original variable (mask_public), with 1 representing that the state has a pubic mask policy, and 0 meaning there is no explicit public mask policy in the state. The output below shows that there are 35 states that enforced a public mask mandate and 16 states that did not at the time the dataset was compiled. This ratio looks reasonable enough for analysis, considering that there is a decent number of samples within each group from the original sample size of 51.

```
df <- df %>%
  mutate(
    mask_public_bool = case_when(
      mask_public == 0 ~ 0,
      !(mask_public == 0) ~ 1
      )
  )

cat('Count of states that enforced a mask mandate in public spaces: ',
    length(df$mask_public_bool[df$mask_public_bool==1]))
```

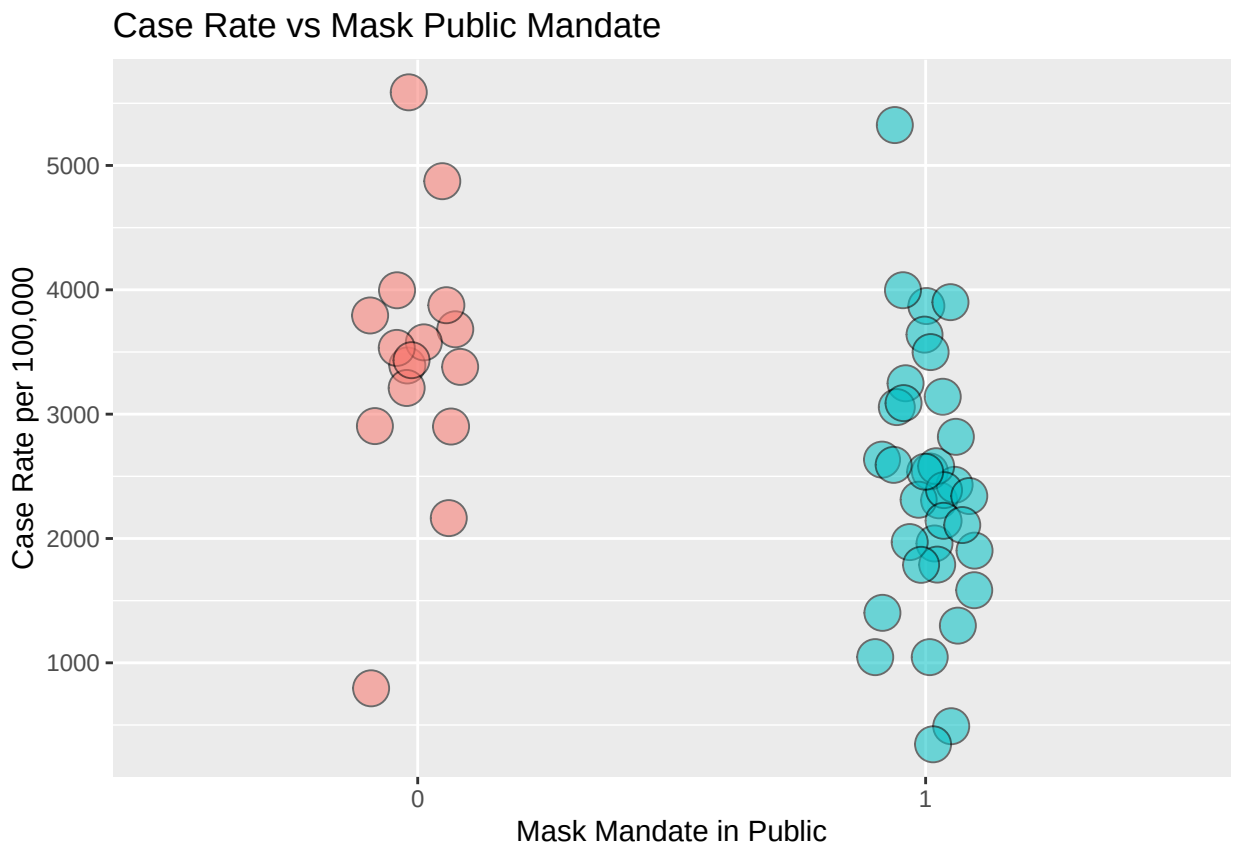
```
## Count of states that enforced a mask mandate in public spaces:  35
```

```
cat('\nCount of states did not enforce a mask mandate in public spaces: ',
    length(df$mask_public_bool[df$mask_public_bool==0]))
```

```
##
## Count of states did not enforce a mask mandate in public spaces:  16
```

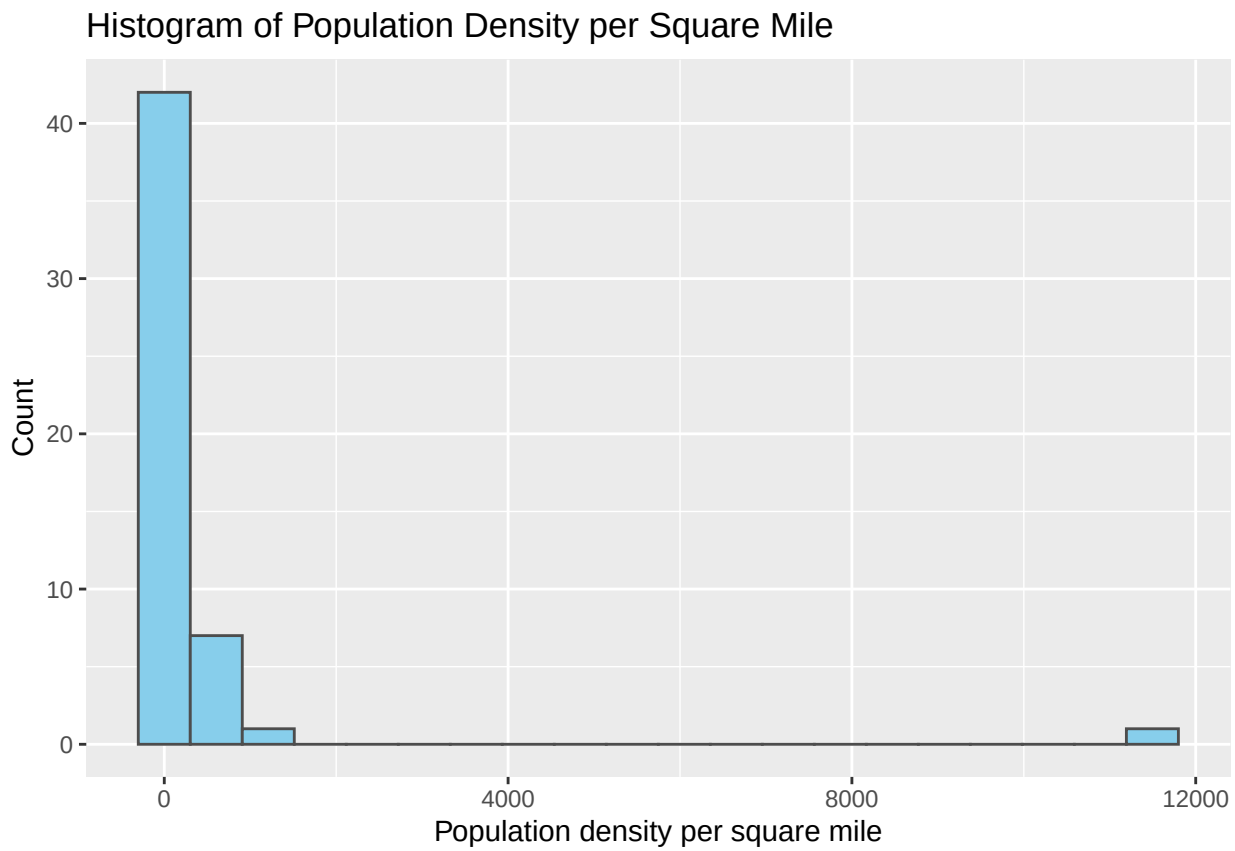
The plot below shows the relationship between the mask policy and case rate. When comparing the states that did not implement a mask policy (value = 0) and those that did (value = 1), it can be seen that the data points for those that have no policy cluster around a higher case rate range, while the data points for the states with a mask policy cluster at a lower case rate range.

```
df %>%
  ggplot(mapping = aes(factor(mask_public_bool), case_rate_100k))+
  geom_jitter(shape=21, size=6, aes( fill = factor(mask_public_bool), alpha=0.5), width=0.1)+
  theme(legend.position = "none") +
  labs(
    title = 'Case Rate vs Mask Public Mandate',
    x = 'Mask Mandate in Public',
    y = 'Case Rate per 100,000'
  ) +
  scale_x_discrete(labels=c("0","1"))
```



Next, the distribution of the population density variable is examined.

```
# Plot the distribution in a histogram
histogram_of_pdensity <- df %>%
  ggplot(aes(x = population_density)) +
  geom_histogram(fill = 'skyblue', color = 'grey30', bins = 20) +
  labs(
    title = 'Histogram of Population Density per Square Mile',
    x = 'Population density per square mile', y = 'Count')

histogram_of_pdensity
```



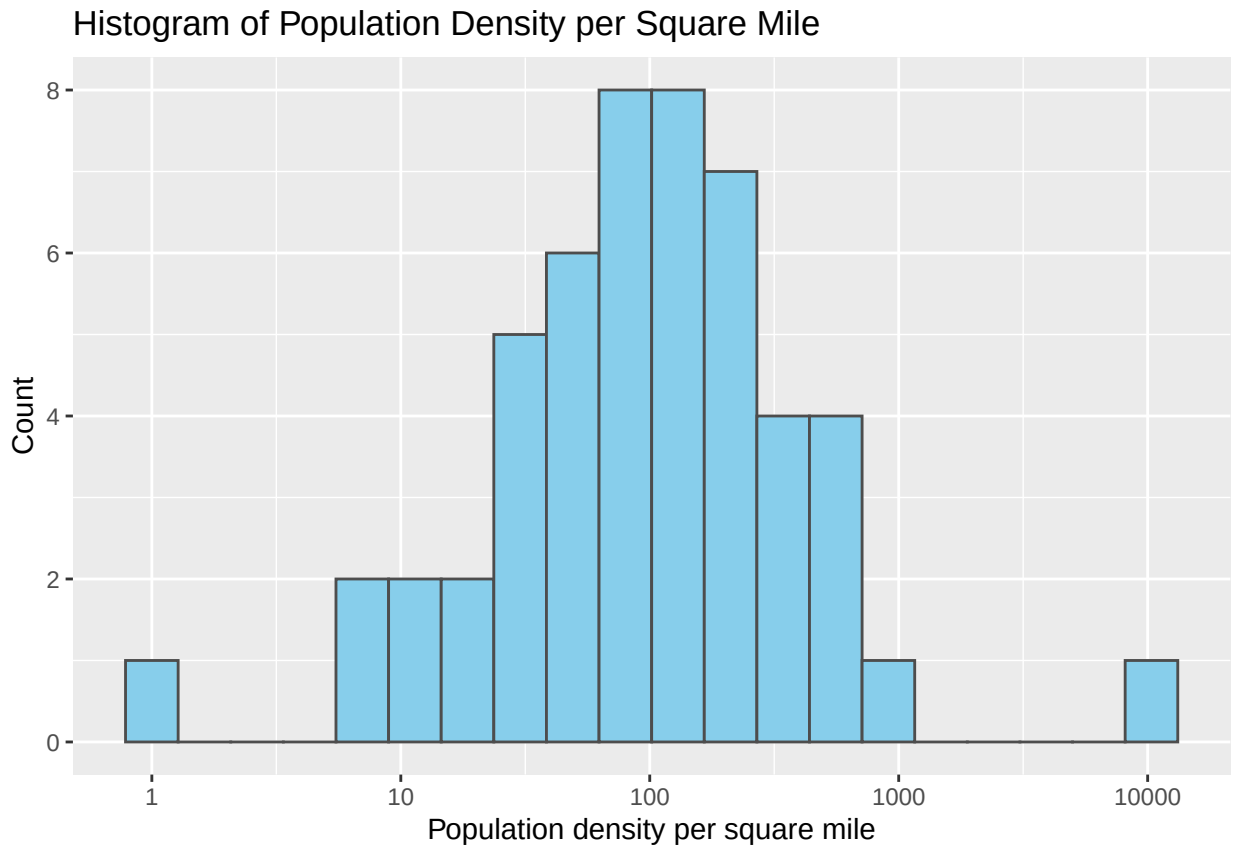
As can be seen from the histogram, although most of the population density clusters in the 0 to 1500 range, there is some grouping of outliers that are very far from this concentration. When an analysis is performed, it can be seen that there is only one data sample outlier, which is D.C. with a value of 11,496. This is given due to the fact that D.C. is a district that solely consists of a large city. Given that this causes the data to be skewed, the logarithm is taken to scale the variable. Alternatively, the data point could have been dropped from the sample, but given the already small sample size, it was determined that a better approach would be to keep it as a data point. Once this transformation was performed, it is shown that there is a relatively normal distribution of population densities (see figure below).

```
# Find the outlier data points
outliers <- subset(df, population_density > 4000)
paste(outliers$State, '=', outliers$population_density)
```

```
## [1] "District of Columbia = 11496.81"
```

```
# Transform the variable by taking the logarithm and assign it to a new variable
df <- df %>%
  mutate(l_population_density = log10(population_density))

# Plot the new distribution in a histogram
histogram_of_pdensity <- df %>%
  ggplot(aes(x = l_population_density)) +
  geom_histogram(fill = 'skyblue', color = 'grey30', bins = 20) +
  labs(
    title = 'Histogram of Population Density per Square Mile',
    x = 'Population density per square mile', y = 'Count') +
```

```
  scale_x_continuous(breaks = seq(0, 5, 1), labels = 10^(seq(0,5,1)))

histogram_of_pdensity
```



2-3-2. Model 3 Regression

Next, the regression model for model 3 is created based on the existing work from model 2. Two additional variables discussed in model 3 will be introduced given the maximalist approach for model 3. The first variable is the "mask_public_bool" (whether the state enforced a public mask mandate or not) that represents the policy. The second variable introduced is "population_density" (the variable is transformed by the natural logarithm as discussed earlier) which is considered an important factor related to the spread of disease.

Based on the results of the data exploration above, the following equation is used to generate the regression analysis for model 3.

$$case\_rate\_100k = \beta_0 + \beta_1 senior\_per\_100k + \beta_2 poverty\_per\_100k + \beta_3 unemployed\_per\_100k + \beta_4 mask\_public\_bool +$$

The regression coefficient shows that mask_public_bool variable is highly significant. The result also suggests that the public mask mandate policy has a negative relationship with the case rate. Specifically, the COVID-19 case rate decreases by 945 out of every 100,000 people if the state enforces a public mask mandate, holding all else constant.

Conversely, the population density variable is not significant in the regression even though it is considered as a major factor in stymieing the spread of the virus. According to the regression, there is a failure to reject the

null hypothesis that population density is not related to the COVID-19 case rate. There could be multiple factors causing the population density to be insignificant. For example, the population density is calculated by population divided by total square miles. As a result, the numbers can be diluted for states that have major cities with high population density, but very little urbanization in the majority of the rural area. In addition, as stated earlier in subsection 2-3-1 of this report, D.C. is considered as a state despite the fact that the demographic and geographical characteristics differ from other states. And due to the small sample size, individual variables such as this that exhibit unique characteristics could lead to a certain degree of bias.

It was decided that the "population_density" variable should be kept in model 3 despite its insignificance for several reasons. First, adding population density variable has little effect on the coefficients of the control variables in model 2, which helps to demonstrate the robustness of the results. Second, there is no evidence that the variable effects the validation of the CLM assumptions compared to those in model 2. Finally, the variable offers meaningful information that is relevant to the main research question.

```
model3 <- lm(case_rate_100k ~ senior_per_100k
              + poverty_per_100k
              + unemployed_per_100k
             + mask_public_bool
             + log(population_density), data = df)
coeftest(model3)
```

```
##
## t test of coefficients:
##
##                            Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)             6146.213024 1273.196469  4.8274 1.631e-05 ***
## senior_per_100k           -0.241854    0.063646 -3.8000  0.000432 ***
## poverty_per_100k           0.196709    0.061312  3.2083  0.002462 **
## unemployed_per_100k       -0.364264    0.179874 -2.0251  0.048813 *
## mask_public_bool        -945.637645  297.728931 -3.1762  0.002695 **
## log(population_density)   93.456585   96.285338  0.9706  0.336925
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

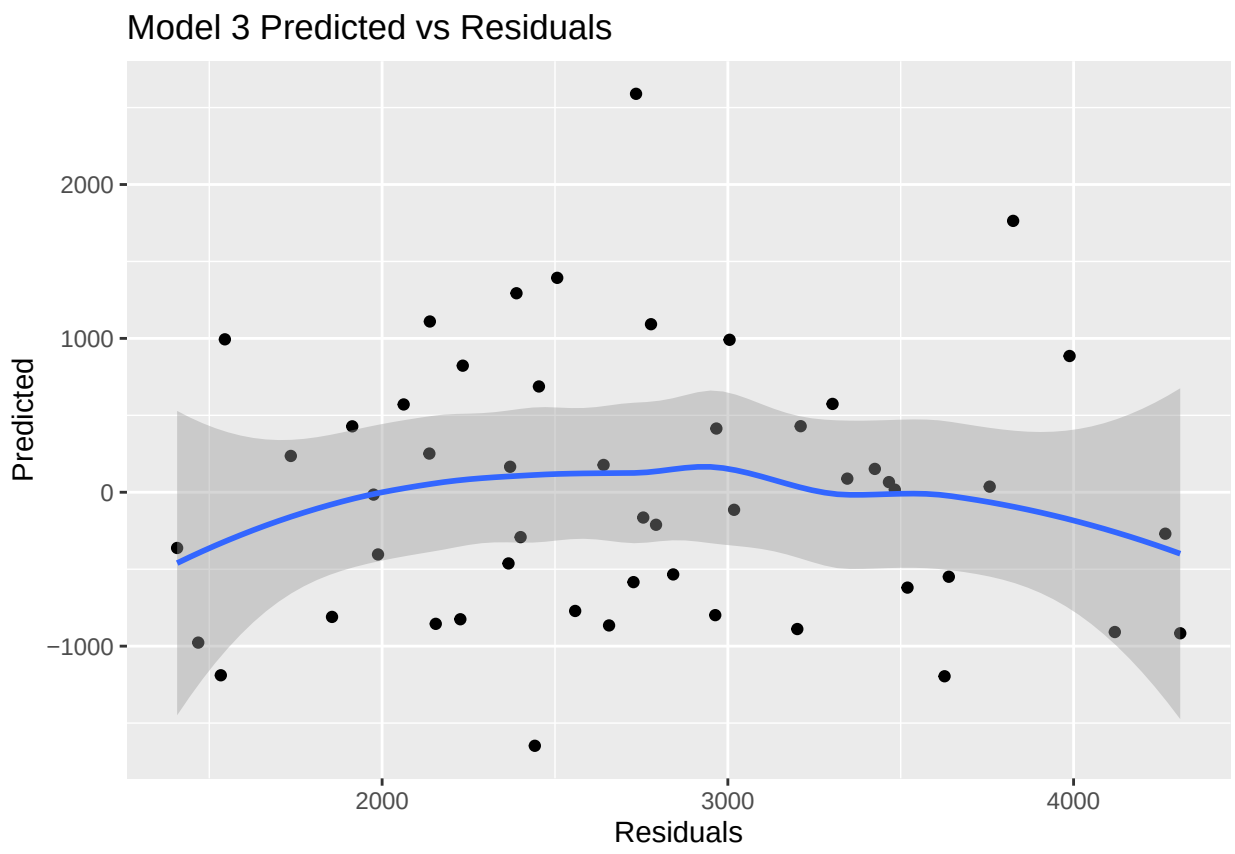
2-3-3. Model 3 Limitations

i. IID Assumption The first assumption of IID is not detailed within this section given that it is evaluated more generally for all three models within section 3.

ii. Linear Conditional Expectation The residual versus fitted plot shows an almost flat line that is close to zero, with the tails slightly skewed. This implies that there are no strong linear relationships that are not captured beyond the current control variables.

```
df<- df%>%
  mutate(
    model3_preds = predict(model3),
    model3_resids = resid(model3)
  )
df %>%
  ggplot(aes(model3_preds, model3_resids)) +
  geom_point() +
```

```
  stat_smooth() +
  labs(
    title = 'Model 3 Predicted vs Residuals',
    x = 'Residuals',
    y = 'Predicted'
  )
```



iii. No Perfect Collinearity First, the list of variable coefficients shows that no variables were dropped, which means there was no perfect collinearity detected by the R function.

```
model3$coefficients
```

```
##             (Intercept)         senior_per_100k        poverty_per_100k
##           6146.2130242             -0.2418541              0.1967093
##     unemployed_per_100k        mask_public_bool log(population_density)
##              -0.3642643           -945.6376446             93.4565854
```

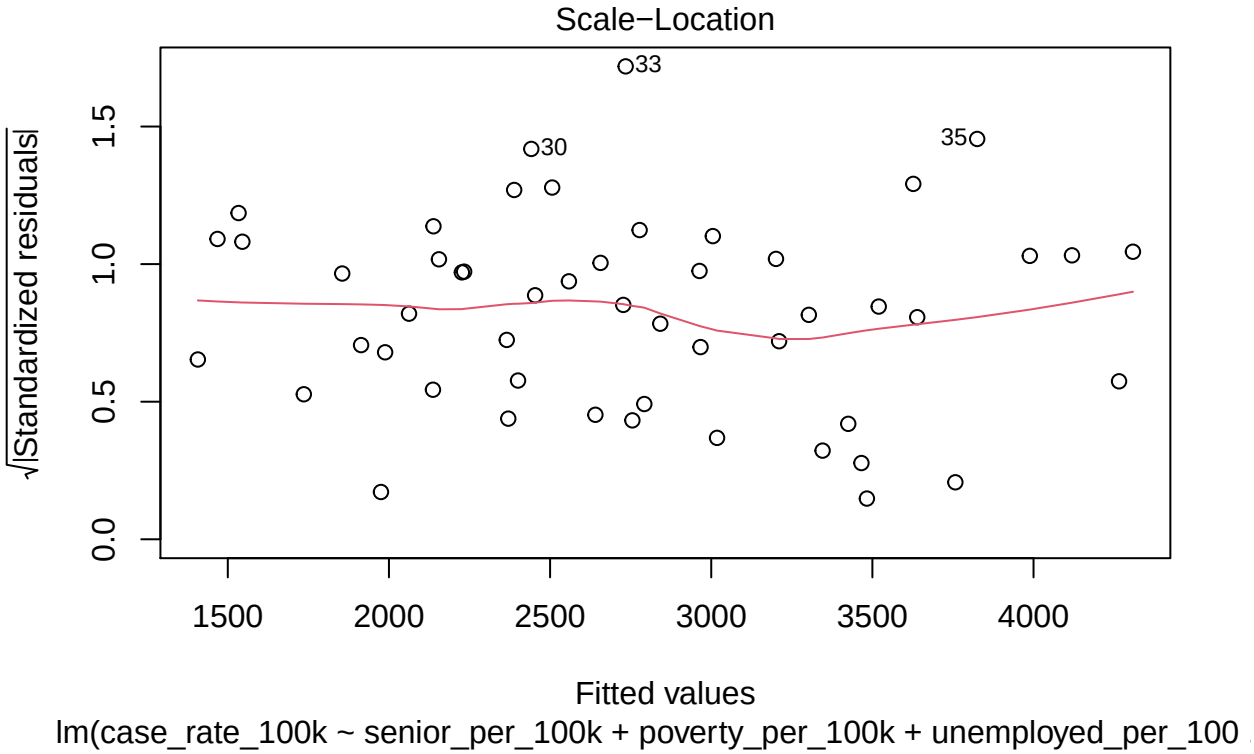
Next, the variance inflation function was run in order to check for strong collinearity. Upon examination of the ouptut below, it can be seen that all the variables have variance inflation factors of less than 4. As a result, there is no strong evidence of collinearity within the model.

```
vif(model3)
```

```
##         senior_per_100k        poverty_per_100k     unemployed_per_100k
##                1.079612                1.882855                2.268679
##        mask_public_bool log(population_density)
##                1.208219                1.282422
```

iv. Homoskedastic Errors According to the scale-location plot, the standardized residuals fall into a reasonable range and the curve looks smooth without strong deviations. There is no significant evidence that the model violates the homoskedasticity assumption. For reference, the marked data points are New Hampshire (30), New York (33), and North Dakota (35).

```
plot(model3, which=3)
```



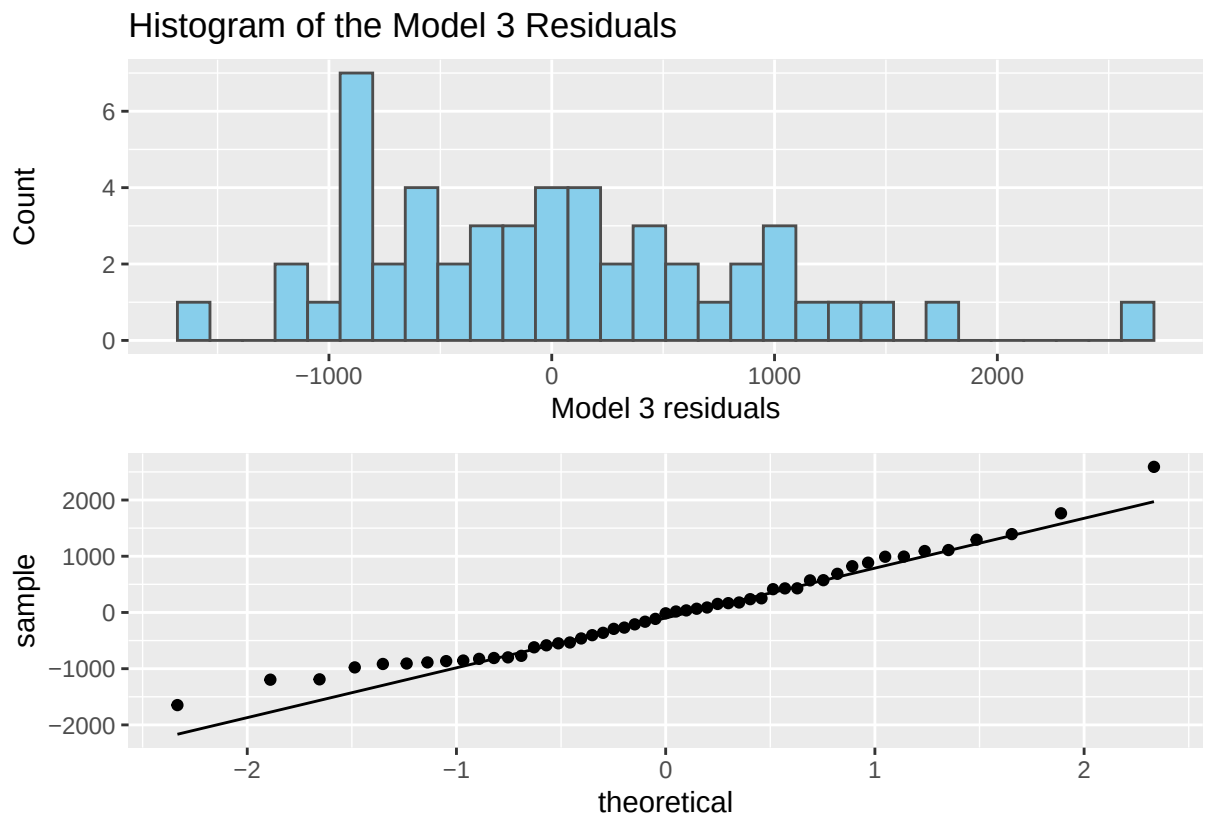
Additionally, a Breusch-Pagan test is run to check the level of heteroskedasticity. The test result shows it fails to reject the null hypothesis, which indicates that there is no strong evidence of heteroskedasticity in model 3.

```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 1.6389, df = 5, p-value = 0.8965
```

v. Normally Distributed Errors To assess the assumption of normality of error distribution, both the histogram and the Q-Q plot of the residuals are shown as follows.

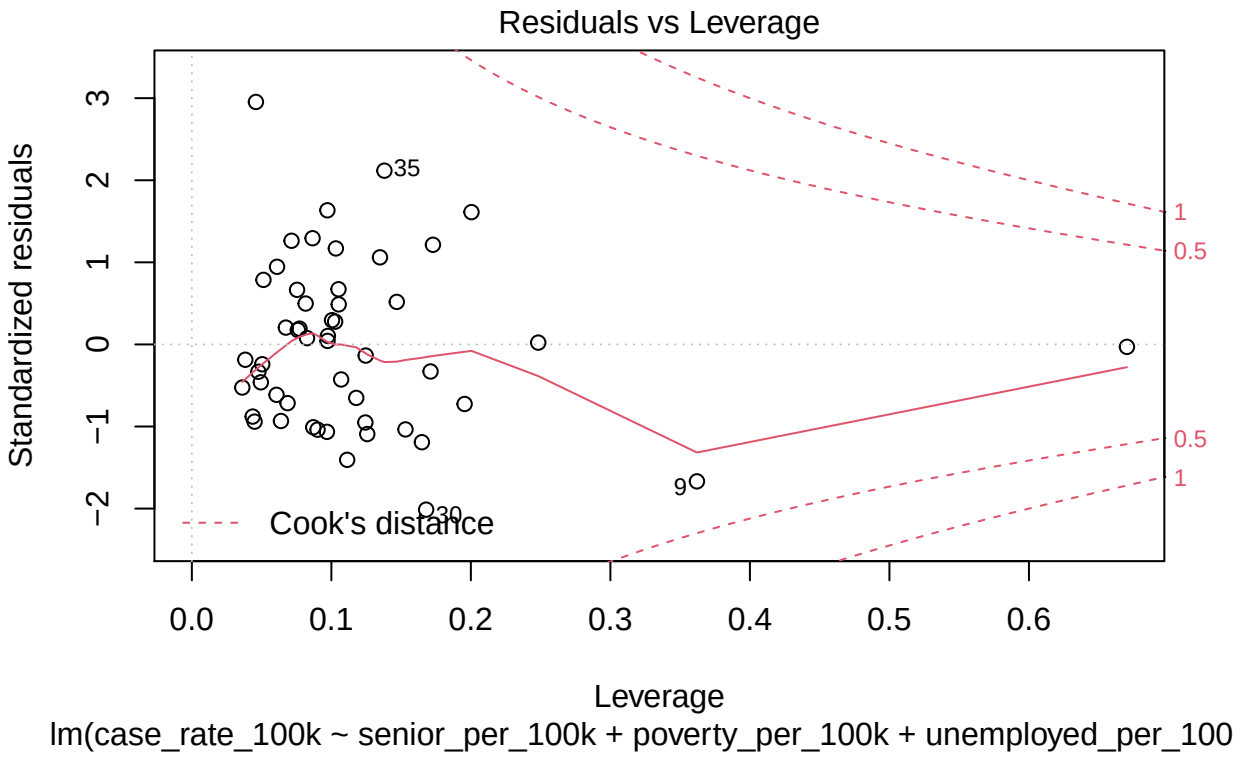
```
plot_one <- df %>%
  ggplot(aes(x = resid(model3))) +
  geom_histogram(fill = 'skyblue', color = 'grey30', bins=30) +
  labs(title =  "Histogram of the Model 3 Residuals",
       x = "Model 3 residuals", y = 'Count')

plot_two <- df %>%
  ggplot(aes(sample = resid(model3))) +
  stat_qq() + stat_qq_line()

plot_one / plot_two
```



There is a fairly normal distribution that can be seen from the histogram above. In the Q-Q plot, the points generally stay close to the normal line, with the tails exhibiting a certain degree of skew. However, in general, it implies no deviation from normality. Overall, the normally distributed errors assumption is satisfied for model 3.

vi. Influence of Data Points From the Cook's distance plot, there is no strong evidence of concerning outliers. For reference, the marked data points are D.C. (9), New Hampshire (30), and North Dakota (35).

```
plot(model3, which=5)
```



3. Limitations of the Models

The most general requirement for all the linear models is that data points be independent and identically distributed (IID). While for the most part the events of one state will primarily affect that single state, travel by individuals spreading the virus across state lines is an inevitable occurrence. This may cause the spread of a virus within one state to be influenced by those around it, or generally by its geographical location within the country. In addition, all states are involved in the same market economy system to varying degrees, even though there are economic performance divisions across the states. Likewise, certain regions of the country may also have similar demographics due to a shared history across general geographic regions. While this is acknowledged to occur, the effect is likely minor compared to the policy data collected state by state which would affect the spread of the virus. Therefore it is relatively safe to reason that this assumption has been met.

The identical distribution of states as individual samples in the model is also questionable. Similar to how connections or commonalities between states make them less than truly independent, certain factors would also influence the distribution so that some states are representative of a different underlying distribution compared to others. For example, if one geographic region as a whole has a factor which influences the spread of the virus differently than another, multiple underlying distributions may be present in the sample. If still a new geographic region with many states was added to the sample, and the measures of central tendency of the data continued to evolve, there would be a further indication that the underlying distribution is not identical. Without any possibility of refining the data sample however, as states would be immutable in this regard, the regression analysis proceeds with the recognition of possible violations in the IID assumption.

In examining the results of various diagnostics across different models, the CLM assumptions are generally better met as variables continue to be added. In the first model, deviations away from certain assumptions

are the strongest. This is seen mostly with regards to the linear conditional expectation assumption, where the residuals against the fitted values have a noticeable inverted "V" shape as opposed to being constant around zero. In the second model, the same diagnostic shows that the assumption is better met. This is further improved upon in the third model, where local variations in the linear conditional expectation diagnostic plot are generally smoother and more centered around zero. The assumption with the greatest exception to this is that of no high collinearity within the models, as each coefficient in each model was shown to have VIF values far below the cut off value of four.

However, when adding additional variables in the third model, other assumptions begin to be less well met. For example, the normality of the error term in the model is less well fit than it was in the earlier models. It can be seen in the third model's Q-Q plot that for fitted values at the edges of the model's range, there exists the greatest deviation away from normality. This assumption is generally less important than that of the linear conditional expectation, however, which is a fundamental part of the CLM. Therefore, the third model can still be considered as generally improving the overall analysis.

In addition, with more variables in the third model, outliers appear to have more of an impact than they did previously. This is likely due to the limited sample size within the models. With less degrees of freedom in the third model, certain points can be seen to more closely approaching the boundaries of the Cook's distance plot, although they are still not extreme enough to be considered outliers. Increasing the resolution of the data, such as examining COVID-19 cases at the county or city/town level, would help to reduce this by increasing the overall degrees of freedom.

Setting aside the uncertainty provided that the sample is qualified for the IID condition, all three of the models generally meet the CLM assumptions. Therefore, the coefficients generated from the regression analyses are unbiased estimates of the actual relationship investigated in this report. The uncertainty associated with these coefficients is also unbiased. Moreover, it is plausible to apply t-tests based on classical standard errors for these regression coefficients. Thus, the discussions and conclusions drawn from these regression results are assumed to be credible. However, if the dataset was more granular (e.g., case rate at the county-level rather than at the state-level), the IID assumption would be better satisfied. It would lead to greater independence for each data point relative to the rest of the data set. It would also take more advantage of the unbiased estimators so far demonstrated in the model, as a higher number of data points would produce more accurate values for the model coefficients. Overall, improving the source data would lead to greater reliability in the regression work and conclusions.

4. Regression Table

With the analysis for the selected variables, the three corresponding models, and the respective diagnostics of the CLM assumptions complete, the models can then be included for comparison in a regression table. Model 1, model 2, and model 3 are included in the regression table below,.

```
se.model1 = coeftest(model1)[ , "Std. Error"]
se.model2 = coeftest(model2)[ , "Std. Error"]
se.model3 = coeftest(model3)[ , "Std. Error"]

stargazer(model1, model2, model3, type = "text",
          se = list(se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = paste("Table 1: The effect of population demographics and mask",
          "policy on COVID-19 case rate"))
```

```
##
## Table 1: The effect of population demographics and mask policy on COVID-19 case rate
## ================================================================================================
```

```
##                                        Dependent variable:
##                          ---------------------------------------------------------
##                                              case_rate_100k
##                                (1)                 (2)                 (3)
## ------------------------------------------------------------------------------------
## senior_per_100k             -0.174**            -0.231**            -0.242***
##                              (0.064)             (0.069)             (0.064)
##
## poverty_per_100k                                 0.232***            0.197**
##                                                  (0.062)             (0.061)
##
## unemployed_per_100k                              -0.486**            -0.364*
##                                                  (0.168)             (0.180)
##
## mask_public_bool                                                     -945.638**
##                                                                      (297.729)
##
## log(population_density)                                               93.457
##                                                                      (96.285)
##
## Constant                   5,613.667***        5,869.330***        6,146.213***
##                             (1,273.196)         (1,362.851)         (1,273.196)
##
## ------------------------------------------------------------------------------------
## Observations                    51                  51                  51
## R2                             0.101               0.311               0.439
## Adjusted R2                    0.082               0.267               0.376
## Residual Std. Error     1,088.808 (df = 49)  973.292 (df = 47)   897.547 (df = 45)
## F Statistic             5.477* (df = 1; 49) 7.059*** (df = 3; 47) 7.034*** (df = 5; 45)
## ====================================================================================
## Note:                                                  *p<0.05; **p<0.01; ***p<0.001
```

Initially addressing the progression of the models at a macro level, the regression table highlights the fact that as the models build upon one another, they seem to get better at explaining the relationship between the input and output variables. Between model 1, 2, and 3, both the R2 and stricter adjusted R2 value increases as variables are added. The R2 value reflects the amount of variation that can be explained in the model on a 0-1 scale, while the adjusted R2 value does the same but also takes into account the number of variables measured (as can be seen in the decreasing degrees of freedom, or df). The adjusted R2 value jumps from a value of 0.082 in model 1 to 0.267 in model 2, which is more than a 200% increase. The value again increases, albeit by a smaller amount, from model 2 to a value of 0.376 in model 3, which represents a 40% jump. When the R2 value is not adjusted, this value increases to 0.439 for model 3. Although not extraordinarily high given the scale, the value shows that that model 3 still explains much of the variance and is a relatively good predictor of the data, especially when compared to model 1. An additional value that reflects this improvement is the fact that the F-statistic is significant for all three models. The F-statistic is used to test whether or not to use the full model (inclusive of all input variables) or a reduced one with fewer variables. Given that the null hypothesis for this test is that the reduced model should be used over the full model, this can be safely rejected in favor of the full model. Additionally, it can be seen that the standard errors for the significant variables are relatively small and are consistent across the models.

When diving into each model more closely and investigating the individual variable coefficients themselves, there is further evidence for robustness in the results across the models.

In model 1, the coefficient of -0.174 for the senior rate is significant with a p-value of 0.0234 (per the regression results from subsection 2-1-2). When the covariate variables of unemployment rate and poverty rate are added in to model 2, the coefficient for senior rate adjusts to a greater magnitude of -0.231 while the p-value

appreciably decreases to a value of 0.0015 (per the regression results from subsection 2-2-2). This explains that not only has the magnitude of the relationship between senior rate and case rate increased, but that the relationship is also more statistically significant. In addition, the coefficients of the covariate variables themselves are also highly significant with p-values of 0.0005 and 0.0059 for poverty rate and unemployment rate respectively. The magnitude of the coefficient of poverty rate is almost equivalent to that of the senior rate, with a coefficient of 0.232, while the unemployment rate has stronger negative coefficient of -0.486 (roughly twice the magnitude of the other two variables).

Moving onto model 3, the magnitude of the coefficient for senior rate increases once again to a value of -0.242, while the statistical significance increases with a decreased p-value of 0.0004 (per the regression results from the subsection 2-3-2). As highlighted before, the more complex model does a better job of predicting the effect of the senior rate variable. Conversely, the p-values for the poverty rate and unemployment rate coefficients become less significant, as they increase to 0.002 and 0.049 respectively. There are also detectable changes in the coefficient values, as the magnitude for both decreases to 0.197 and -0.364 respectively. The changes in the p-values are likely a direct result of the added variables of mask policy and the population density, where these variables introduce some additional variance in the model 2 covariates. Notably, the coefficient for the mask policy covariate is significant with a p-value of 0.0027 and has by far the largest magnitude of coefficient, with a value of -945.6. The population density covariate is not significant with a p-value of 0.3369.

In terms of practical significance, there is great discrepancy between the coefficient of the mask policy variable compared to the other three significant features. With a coefficient of -945.6, this means that given the binary mask policy variable, when a mask policy is implemented (value of 1), there is an decrease in case rate of roughly 945 per 100,000 compared to when the policy is not in place. The next largest coefficient magnitude is from the unemployment rate variable, where in the final model, the coefficient of -0.364 translates to a decrease of roughly 1 case per 100,000 with every 3 additional people unemployed per 100,000. Similarly, the coefficients of both poverty rate and senior rate are relatively small. For every 5 additional people per 100,000 who are below the poverty line, there is roughly 1 additional case per 100,000, and for every 4 additional seniors per 100,000 within the state population, there is roughly 1 fewer case per 100,000.

5. Omitted Variables

Omitted variables from the model may affect both the internal specifications of included variables as well as the general applicability of the model in explaining the output variable. For this analysis, the omitted variables mentioned are hypothetical. Therefore, without the data available to conduct a follow-up analysis on the omitted variables, only the direction of the bias may be estimated, as opposed to the actual effect size.

5-1. Urban vs Rural Population Ratio

One of the variables that was hypothesized as being potentially influential is a ratio of a state's urban to rural population. Currently, the dataset contains population density information as averaged across each state as a whole. While population density is potentially useful, it was found to not be significant in model 3. This may be due to states having large urban centers, where population density would be high, that lose their impact due to the state also containing large areas of land that is sparsely populated. One example of this would be New York, which contains several of the most densely populated counties in the country, but also has large rural areas.

An urban to rural ratio would be explanatory of the population density variable as a whole, where a higher urban to rural ratio would result in a higher statewide population density generally. With a greater fraction of urban residents, people would also be less likely to be able to social distance as effectively, increasing the case rate. Therefore, the coefficient for population density in the original model would be larger and more positive than a model where this ratio is included, resulting in a bias that is positive and away from zero.

Assuming population density is an effective proxy for an urban to rural ratio variable, however, the original effect described in the models can be assumed to be real.

5-2. Median Income

Knowledge of the median income for a state would give an indication of the types of jobs and earnings available to the labor force within that state. It could be relevant to case rate to understand how many people work in low paying jobs and thus may more frequently be in close contact with others (such as the service industry, gig industry, etc.). Higher median income may imply that more workers are able to function remotely with greater social distancing. A higher median income would then be negatively related to the COVID case rate. It would also be explanatory in a way that makes the poverty percentage variable less impactful, where either of these variables may be considered as proxies for one another. Because inclusion of the median income variable would lead to a lower magnitude poverty percentage variable, and the coefficient of the poverty percentage variable is positive, this would cause a positive omitted variable bias away from zero. Similar to the first omitted variable, however, poverty percentage may also be assumed to be a proxy for median income in that they both describe the economic state of a state and what that implies for its population. Because of this, the original effect in the model can be assumed to be real.

5-3. Gender Ratio

The ratio of gender between women and men may impact the COVID case rate in that, according to external studies, men are more susceptible to viral infections and generally have lower life expectancy and are less healthy as a result of lifestyle differences. Greater susceptibility may relate to the case rate in that men may more often become infected when exposed to COVID under similar circumstances as women. Poorer health may also impact the case rate in that a greater proportion of the COVID cases among men may be symptomatic, with fewer occurrences of asymptomatic cases that go undetected. In either of these cases, a higher fraction of men among the population would lead to a higher COVID case rate overall.

This variable could also be partially explained by the senior percentage variable, because of the increase of women in gender ratios generally as people get older. Therefore, introducing the gender ratio variable into the model would lower the value of the coefficient to the senior percentage variable to be less negative. This would be a negative omitted variable bias towards zero. With a negative bias, and no existing variable in the dataset serving as a good proxy for gender ratio, the effects in the original model can be assumed to be real.

However, the magnitude of this change is likely to not be large due to other causal mechanisms between seniors and COVID case rate. While men may more frequently present COVID symptoms, older people in general are also more susceptible to the virus due to their physical condition and would also present more symptoms than non-seniors. When comparing the two, seniors are generally considered to be more at-risk of COVID complications than men across all ages, and so the senior percentage would still have a greater impact on the overall COVID case rate relative to gender ratio.

5-4. Mask Use Among Population

Currently, the model incorporates state policy around masks, with a variable indicating whether or not a state mandated a mask policy for public spaces. While this is useful for describing an action a state government can take against case rate, it does not fully capture the compliance with this policy among the population. Instead, a variable which would indicate the fraction of the population that actively wears masks would better capture the mechanism of how the virus is transmitted among people.

A variable capturing the real mask use among the population would likely reduce the explanatory power of the state mask mandate variable to be less negative. This would cause the bias of the omitted variable to be negative and towards from zero. It is also reasonable to assume that state mask policy may be an effective

enough proxy for real mask use among the population, in which case the original effect in the model may be assumed to be real.

5-5. GDP Percentage of Tourism

The economic standing of a state not only affects the individual households of its population, but also includes how it interacts with other states and how its peoples' lives might have been economically changed by the pandemic. One way that a state may increase its number of COVID cases is both from tourists or others visiting during the pandemic and bringing disease, but also if its own population must work in close contact with others serving them in the tourism industry. In addition, if certain key industries to a state, such as tourism, laid off many workers, they would likely have to take lower paying close-contact jobs such as in the service industry or gig economy in order to continue to have a livelihood. This would further lead to a higher COVID case rate.

These effects may be included with a variable that demonstrates how much of a state's GDP is from tourism. It would be related to other economic indicator variables, such as the "real time" poverty rate during the pandemic. Although poverty is a broader phenomena, the number of people in poverty in the pandemic could be higher due to negative growth of the tourism industry. Therefore, it is expected that states with a greater focus on tourism may also have more people in poverty in the pandemic. While poverty rate is included in the model, it is based off of data from 2018 before any pandemic related effects would have occurred. Assuming real time poverty was included in the model, a GDP percentage of tourism would take some of the explanatory power of the poverty variable away, resulting in its model coefficient becoming less positive. This would result in a positive omitted variable bias away from zero. Without a clear proxy in the dataset, and with the nature of this variable touching on the IID assumption between interaction among states, it should be further investigated whether a model including only poverty may be affected by omitted variable bias. Again, because the model in this analysis uses 2018 poverty data, this would not be related to a bias from this omitted variable.

6. Conclusion

The goal of the analysis was to address the research question of how the COVID-19 case rate is related to the distribution of population demographics and mask-related policy of a state. In order to address this question, the analysis was broken up into three phases of investigation with corresponding linear models.

- Phase 1: To investigate the direct relationship between COVID-19 case rate and the senior rate within a state
- Phase 2: To build upon Phase 1 by including two additional variables of population demographics, poverty rate and unemployment rate, and investigating the overall relationship on COVID-19 case rate
- Phase 3: To build upon Phase 2 by adding a variable to measure the relationship of the implementation of a mask-policy on COVID-19 case rate. A variable for population density was also added as an additional measure of population demographics

From the three phases of investigation, there was strong evidence in the data to suggest that senior rate, unemployment rate, and mask policy are all statistically significant in relation to case rate and have a negative relationship with case rate, while poverty rate has a positive relationship. The data for population density was not statistically significant enough to suggest a relationship.

Starting with the variable with the largest coefficient magnitude, it can be seen that with the implementation of a mask policy in a state there are 945 fewer cases per 100,000 compared to a state without any implementation of such policy. Given that research from leading health organizations shows that the wearing of masks helps to decrease the chance of the transmission of COVID-19, it could be suggested that a mask policy is effective in increasing the adoption of wearing masks among the state population. Next, it can be seen that

there is a decrease of roughly 1 case per 100,000 with every 3 additional people unemployed per 100,000. This suggests that as unemployment increases, the case rate decreases, which could be a result of the fact that those who are unemployed may be less mobile as they are not commuting into work, lowering the likelihood of contracting the virus. Following unemployment, when examining the poverty variable there is actually an increase of roughly 1 case per 100,000 for every 5 additional people per 100,000 living under the federal poverty line. Although there is some intuition behind the unemployment and poverty rates having similar relationships with case rate, it can be seen from the test of collinearity in subsection 2-2-3 that these two variables are not explaining the same effects. This difference is likely due to the fact that those living below the poverty line have less access to suitable living conditions and sanitary goods that help to alleviate the possibility of COVID-19 infection. As a caveat, it should be noted that these the poverty and unemployment figures are from a 2018 survey, so it does not reflect the full picture of the state demographics at the time of when the COVID-19 related data was captured. Finally, the coefficient of the senior rate is negative, with an additional 4 seniors per 100,000 translating to roughly 1 fewer case per 100,000. As touched upon in the introduction, this could be largely down to the fact that it is widely broadcasted that seniors fall within the high risk category in terms of having severe complications from COVID-19, and therefore they are more diligent in the precautions they take against contracting the virus.

In conclusion, going back to our original question, it can be suggested within the data that the effect of a mask-related policy is something that should not be neglected in comparison to the population demographics. Therefore, it seems that the COVID-19 case rate at the state-level is dependent on masks, and that ultimately, there are effective policy and control measures that can be taken by state governments that have some contribution towards combating its spread relative to factors that cannot be controlled. However, it should be noted that these insights are qualified by the assumptions and limitations laid out throughout the report, namely that there are potential issues in meeting the IID assumption and that there is explanatory information that is lost from both the variables investigated as well as the those that were not included in the dataset.

References

Centers for Disease Control and Prevention. 2020. "Race, Ethnicity, and Age Trends in Persons Who Died from Covid-19 — United States, May–August 2020." Available at https://www.cdc.gov/mmwr/volumes/69/wr/mm6942e1.htm?s_cid=mm6942e1_e&ACSTrackingID=USCDC_921-DM40574&ACSTrackingLabel=MMWR%20Early%20Release%20-%20Vol.%2069%2C%20October%2016%2C%202020&deliveryName=USCDC_921-DM40574 (2020/12/05).

Mark Mather, and Lillian Kilduff. 2020. "The U.S. Population Is Growing Older, and the Gender Gap in Life Expectancy Is Narrowing." Available at https://www.prb.org/the-u-s-population-is-growing-older-and-the-gender-gap-in-life-expectancy-is-narrowing/ (2020/12/05).